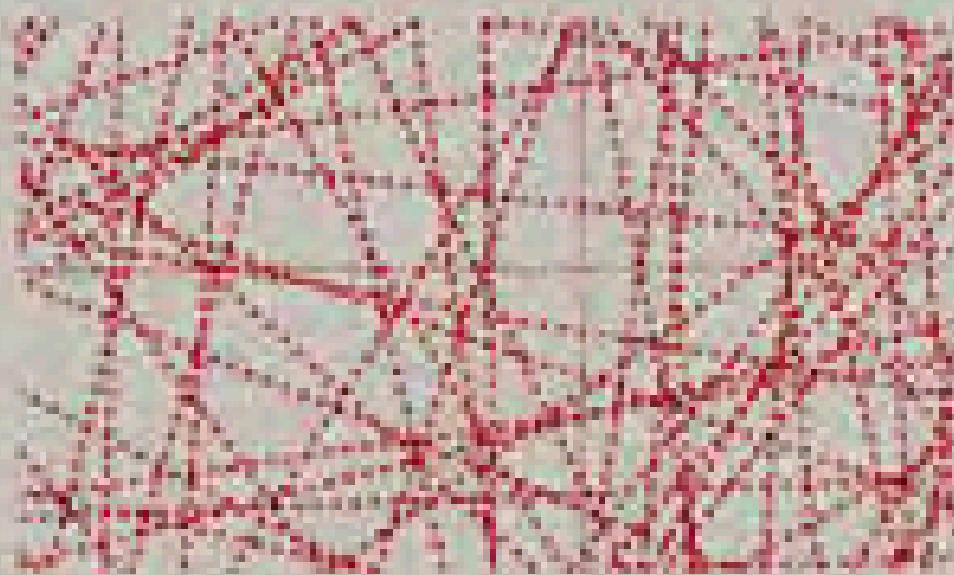


Statistical Learning Theory



Vladimir N. Vapnik

A Williams & Wilkins Book in Adaptive and Learning Systems
for Signal Processing, Communications, and Control
Series Editors: Robert歇夫, Robert歇夫

Adaptive and Learning Systems for Signal Processing, Communications, and Control

Editor: Simon Haykin

Werbos / THE ROOTS OF BACKPROPAGATION: From Ordered Derivatives to Neural Networks and political Forecasting

Krstic *, Kanellakopoulos, and Kokotovic^{*} / NONLINEAR AND ADAPTIVE CONTROL DESIGN

Nikias and Shao / SIGNAL PROCESSING WITH ALPHA-STABLE DISTRIBUTIONS AND APPLICATIONS

Diamantaras and Kung / PRINCIPAL COMPONENT NEURAL NETWORKS: THEORY AND APPLICATIONS

Tao and Kokotovic^{*} / ADAPTIVE CONTROL OF SYSTEMS WITH ACTUATOR AND SENSOR NONLINEARITIES

Tsoukalas / FUZZY AND NEURAL APPROACHES IN ENGINEERING Hrycej / NEUROCONTROL: TOWARDS AN INDUSTRIAL CONTROL METHODOLOGY

Beckerman / ADAPTIVE COOPERATIVE SYSTEMS

Cherkassky and Mulier / LEARNING FROM DATA: CONCEPTS, THEORY, AND METHODS

Passino and Burgess / STABILITY ANALYSIS OF DISCRETE EVENT SYSTEMS

Sánchez-Peña and Sznaier / ROBUST SYSTEMS THEORY AND APPLICATIONS

Vapnik / STATISTICAL LEARNING THEORY

Statistical Learning Theory

Vladimir N. Vapnik

AT&T Research Laboratories



A WILEY-INTERSCIENCE PUBLICATION

JOHNWILEY & SONS, INC.

NEW YORK / CHICHESTER / WEINHEIM / BRISBANE / SINGAPORE / TORONTO

Disclaimer:

This book contains characters with diacritics. When the characters can be represented using the ISO 8859-1 character set (<http://www.w3.org/TR/images/latin1.gif>), netLibrary will represent them as they appear in the original text, and most computers will be able to show the full characters correctly. In order to keep the text searchable and readable on most computers, characters with diacritics that are not part of the ISO 8859-1 list will be represented without their diacritical marks.

This book is printed on acid-free paper.*

Copyright © 1998 by John Wiley & Sons, Inc. All rights reserved.

published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4744. Requests to the publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 605 Third Avenue, New York, NY 10158-0012, (212) 850-6011, fax (212) 850-6008, E-Mail: PERMREQ@WILEY.COM.

Library of Congress Cataloging-in-Publication Data:

Vapnik, Vladimir Naumovich

Statistical learning theory / Vladimir N. Vapnik

p. cm.--(Adaptive and learning systems for signal processing, communications, and control)

Includes bibliographical references and index.

ISBN 0-471-03003-1 (cloth : alk. paper)

1. Computational learning theory. I. Title. II. Series.

Q325.7.V38 1998

008.31--dc21

97-37075

CIP

printed in the United States of America

1 0 9 8 7 6 5 4 3 2 1

In memory of my father

CONTENTS

Preface	<u>xxi</u>
Introduction: The Problem of Induction and Statistical Inference	<u>1</u>
0 1 Learning Paradigm in Statistics	<u>1</u>
0 2 Two Approaches to Statistical Inference Particular (Parametric Inference) and General (Nonparametric Inference)	<u>2</u>
0 3 The Paradigm Created by the Parametric Approach	<u>4</u>
0 4 Shortcoming of the Parametric Paradigm	<u>5</u>
0 5 After the Classical Paradigm	<u>6</u>
0 6 The Renaissance	<u>7</u>
0 7 The Generalization of the Glivenko-Cantelli-Kolmogorov Theory	<u>8</u>
0 8 The Structural Risk Minimization Principle	<u>10</u>
0 9 The Main Principle of Inference from a Small Sample Size	<u>11</u>
0 10 What This Book is About	<u>13</u>
I Theory of Learning and Generalization	
1 Two Approaches to the Learning Problem	<u>19</u>
1 1 General Model of Learning from Examples	<u>19</u>
1 2 The Problem of Minimizing the Risk Functional from Empirical Data	<u>21</u>
1 3 The Problem of Pattern Recognition	<u>24</u>
1 4 The Problem of Regression Estimation	<u>26</u>
1 5 Problem of Interpreting Results of Indirect Measuring	<u>28</u>
1 6 The Problem of Density Estimation (the Fisher-Wald Setting)	<u>30</u>

1 7 Induction Principles for Minimizing the Risk Functional on the Basis of Empirical Data	<u>32</u>
1 8 Classical Methods for Solving the Function Estimation Problems	<u>33</u>
1 9 Identification of Stochastic Objects Estimation of the Densities and Conditional Densities	<u>35</u>
1 9 1 Problem of Density Estimation Direct Setting	<u>35</u>
1 9 2 Problem of Conditional Probability Estimation	<u>36</u>
1 9 3 Problem of Conditional Density Estimation	<u>37</u>
1 10 The Problem of Solving an Approximately Determined Integral Equation	<u>38</u>
1 11 Glivenko-Cantelli Theorem	<u>39</u>
1 11 1 Convergence in Probability and Almost Sure Convergence	<u>40</u>
1 11 2 Glivenko-Cantelli Theorem	<u>42</u>
1 11 3 Three Important Statistical Laws	<u>42</u>
1 12 Ill-Posed Problems	<u>44</u>
1 13 The Structure of the Learning Theory	<u>48</u>
Appendix to Chapter 1 Methods for Solving Ill-Posed Problems	<u>51</u>
A1 1 The Problem of Solving an Operator Equation	<u>51</u>
A1 2 Problems Well-Posed in Tikhonov's Sense	<u>53</u>
A1 3 The Regularization Method	<u>54</u>
A1 3 1 Idea of Regularization Method	<u>54</u>
A1 3 2 Main Theorems about the Regularization Method	<u>55</u>
2 Estimation of the Probability Measure and Problem of Learning	<u>59</u>
2 1 Probability Model of a Random Experiment	<u>59</u>
2 2 The Basic Problem of Statistics	<u>61</u>
2 2 1 The Basic Problems of Probability and Statistics	<u>61</u>
2 2 2 Uniform Convergence of Probability Measure Estimates	<u>62</u>
2 3 Conditions for the Uniform Convergence of Estimates to the Unknown Probability Measure	<u>65</u>
2 3 1 Structure of Distribution Function	<u>65</u>
2 3 2 Estimator that Provides Uniform Convergence	<u>68</u>
2 4 Partial Uniform Convergence and Generalization of Glivenko-Cantelli Theorem	<u>69</u>

2.4.1 Definition of Partial Uniform Convergence	<u>69</u>
2.4.2 Generalization of the Glivenko-Cantelli Problem	<u>71</u>
2.5 Minimizing the Risk Functional Under the Condition of Uniform Convergence of Probability Measure Estimates	<u>72</u>
2.6 Minimizing the Risk Functional under the Condition of Partial Uniform Convergence of Probability Measure Estimates	<u>74</u>
2.7 Remarks about Modes of Convergence of the Probability Measure Estimates and Statements of the Learning Problem	<u>77</u>
3 Conditions for Consistency of Empirical Risk Minimization Principle	<u>79</u>
3.1 Classical Definition of Consistency	<u>79</u>
3.2 Definition of Strict (Nontrivial) Consistency	<u>82</u>
3.2.1 Definition of Strict Consistency for the Pattern Recognition and the Regression Estimation Problems	<u>82</u>
3.2.2 Definition of Strict Consistency for the Density Estimation Problem	<u>84</u>
3.3 Empirical Processes	<u>85</u>
3.3.1 Remark on the Law of Large Numbers and Its Generalization	<u>86</u>
3.4 The Key Theorem of Learning Theory (Theorem about Equivalence)	<u>88</u>
3.5 Proof of the Key Theorem	<u>89</u>
3.6 Strict Consistency of the Maximum Likelihood Method	<u>92</u>
3.7 Necessary and Sufficient Conditions for Uniform Convergence of Frequencies to their Probabilities	<u>93</u>
3.7.1 Three Cases of Uniform Convergence	<u>93</u>
3.7.2 Conditions of Uniform Convergence in the Simplest Model	<u>94</u>
3.7.3 Entropy of a Set of Functions	<u>95</u>
3.7.4 Theorem about Uniform Two-Sided Convergence	<u>97</u>
3.8 Necessary and Sufficient Conditions for Uniform Convergence of Means to their Expectations for a Set of Real-Valued Bounded Functions	<u>98</u>
3.8.1 Entropy of a Set of Real-Valued Functions	<u>98</u>
3.8.2 Theorem about Uniform Two-Sided Convergence	<u>99</u>
3.9 Necessary and Sufficient Conditions for Uniform Convergence of Means to their Expectations for Sets of Unbounded Functions	<u>100</u>
3.9.1 Proof of Theorem 3.5	

3 10 Kant's Problem of Demarcation and Popper's Theory of Nonfalsifiability	<u>106</u>
3 11 Theorems about Nonfalsifiability	<u>108</u>
3 11 1 Case of Complete Nonfalsifiability	<u>108</u>
3 11 2 Theorem about Partial Nonfalsifiability	<u>109</u>
3 11 3 Theorem about Potential Nonfalsifiability	<u>110</u>
3 12 Conditions for One-Sided Uniform Convergence and Consistency of the Empirical Risk Minimization Principle	<u>112</u>
3 13 Three Milestones in Learning Theory	<u>118</u>
4 Bounds on the Risk for Indicator Loss Functions	<u>121</u>
4 1 Bounds for the Simplest Model Pessimistic Case	<u>122</u>
4 1 1 The Simplest Model	<u>123</u>
4 1 2 Bounds for the Simplest Model Optimistic Case	<u>125</u>
4 1 3 Bounds for the Simplest Model General Case	<u>127</u>
4 1 4 The Basic Inequalities Pessimistic Case	<u>129</u>
4 1 5 Proof of Theorem 4 1	<u>131</u>
4 1 5 1 The Basic Lemma	<u>131</u>
4 1 5 2 Proof of Basic Lemma	<u>132</u>
4 1 5 3 The Idea of Proving Theorem 4 1	<u>134</u>
4 1 5 4 Proof of Theorem 4 1	<u>135</u>
4 1 6 Basic Inequalities General Case	<u>137</u>
4 1 7 Proof of Theorem 4 2	<u>139</u>
4 1 8 Main Nonconstructive Bounds	<u>144</u>
4 1 9 VC Dimension	<u>145</u>
4 1 9 1 The Structure of the Growth Function	<u>145</u>
4 1 9 2 Constructive Distribution-Free Bounds on Generalization Ability	<u>148</u>
4 1 9 3 Solution of Generalized Glivenko-Cantelli Problem	<u>149</u>
4 1 10 Proof of Theorem 4 3	<u>150</u>
4 1 11 Example of the VC Dimension of the Different Sets of Functions	<u>155</u>
4 1 12 Remarks about the Bounds on the Generalization Ability of Learning Machines	<u>160</u>
4 1 13 Bound on Deviation of Frequencies in Two Half-Samples	<u>163</u>
Appendix to Chapter 4 Lower Bounds on the Risk of the ERM Principle	<u>169</u>
A4 1 Two Strategies in Statistical Inference	<u>169</u>
A4 2 Minimax Loss Strategy for Learning Problems	<u>171</u>

A4 3 Upper Bounds on the Maximal Loss for the Empirical Risk Minimization Principle	<u>173</u>
A3 3 1 Optimistic Case	<u>173</u>
A3 3 2 Pessimistic Case	<u>174</u>
A4 4 Lower Bound for the Minimax Loss Strategy in the Optimistic Case	<u>177</u>
A4 5 Lower Bound for Minimax Loss Strategy for the Pessimistic Case	<u>179</u>
5 Bounds on the Risk for Real-Valued Loss Functions	<u>183</u>
5 1 Bounds for the Simplest Model Pessimistic Case	<u>183</u>
5 2 Concepts of Capacity for the Sets of Real-Valued Functions	<u>186</u>
5 2 1 Nonconstructive Bounds on Generalization for Sets of Real-Valued Functions	<u>186</u>
5 2 2 The Main Idea	<u>188</u>
5 2 3 Concepts of Capacity for the Set of Real-Valued Functions	<u>190</u>
5 3 Bounds for the General Model Pessimistic Case	<u>192</u>
5 4 The Basic Inequality	<u>194</u>
5 4 1 Proof of Theorem 5 2	<u>195</u>
5 5 Bounds for the General Model Universal Case	<u>196</u>
5 5 1 Proof of Theorem 5 3	<u>198</u>
5 6 Bounds for Uniform Relative Convergence	<u>200</u>
5 6 1 Proof of Theorem 5 4 for the Case $p > 2$	<u>201</u>
5 6 2 Proof of Theorem 5 4 for the Case $1 < p \leq 2$	<u>204</u>
5 7 Prior Information for the Risk Minimization Problem in Sets of Unbounded Loss Functions	<u>207</u>
5 8 Bounds on the Risk for Sets of Unbounded Nonnegative Functions	<u>210</u>
5 9 Sample Selection and the Problem of Outliers	<u>214</u>
5 10 The Main Results of the Theory of Bounds	<u>216</u>
6 The Structural Risk Minimization Principle	<u>219</u>
6 1 The Scheme of the Structural Risk Minimization Induction Principle	<u>219</u>
6 1 1 Principle of Structural Risk Minimization	<u>221</u>
6 2 Minimum Description Length and Structural Risk Minimization Inductive Principles	<u>224</u>
6 2 1 The Idea about the Nature of Random Phenomena	<u>224</u>

6.2.2 Minimum Description Length Principle for the Pattern Recognition Problem	<u>224</u>
6.2.3 Bounds for the Minimum Description Length Principle	<u>226</u>
6.2.4 Structural Risk Minimization for the Simplest Model and Minimum Description Length Principle	<u>227</u>
6.2.5 The Shortcoming of the Minimum Description Length Principle	<u>228</u>
6.3 Consistency of the Structural Risk Minimization Principle and Asymptotic Bounds on the Rate of Convergence	<u>229</u>
6.3.1 Proof of the Theorems	<u>232</u>
6.3.2 Discussions and Example	<u>235</u>
6.4 Bounds for the Regression Estimation Problem	<u>237</u>
6.4.1 The Model of Regression Estimation by Series Expansion	<u>238</u>
6.4.2 Proof of Theorem 6.4	<u>241</u>
6.5 The Problem of Approximating Functions	<u>246</u>
6.5.1 Three Theorems of Classical Approximation Theory	<u>248</u>
6.5.2 Curse of Dimensionality in Approximation Theory	<u>251</u>
6.5.3 Problem of Approximation in Learning Theory	<u>252</u>
6.5.4 The VC-Dimension in Approximation Theory	<u>254</u>
6.6 Problem of Local Risk Minimization	<u>257</u>
6.6.1 Local Risk Minimization Model	<u>259</u>
6.6.2 Bounds for the Local Risk Minimization Estimator	<u>262</u>
6.6.3 Proofs of the Theorems	<u>265</u>
6.6.4 Structural Risk Minimization Principle for Local Function Estimation	<u>268</u>
Appendix to Chapter 6. Estimating Functions on the Basis of Indirect Measurements	<u>271</u>
A6.1 Problems of Estimating the Results of Indirect Measurements	<u>271</u>
A6.2 Theorems on Estimating Functions Using Indirect Measurements	<u>273</u>
A6.3 Proofs of the Theorems	<u>276</u>
A6.3.1 Proof of Theorem A6.1	<u>276</u>
A6.3.2 Proof of Theorem A6.2	<u>281</u>
A6.3.3 Proof of Theorem A6.3	<u>283</u>

7 Stochastic Ill-Posed Problems	<u>293</u>
7.1 Stochastic Ill-Posed Problems	<u>293</u>
7.2 Regularization Method for Solving Stochastic Ill-Posed Problems	<u>297</u>
7.3 Proofs of the Theorems	<u>299</u>
7.3.1 Proof of Theorem 7.1	<u>299</u>
7.3.2 Proof of Theorem 7.2	<u>302</u>
7.3.3 Proof of Theorem 7.3	<u>303</u>
7.4 Conditions for Consistency of the Methods of Density Estimation	<u>305</u>
7.5 Nonparametric Estimators of Density Estimators Based on Approximations of the Distribution Function by an Empirical Distribution Function	<u>308</u>
7.5.1 The Parzen Estimators	<u>308</u>
7.5.2 Projection Estimators	<u>313</u>
7.5.3 Spline Estimate of the Density Approximation by Splines of the Odd Order	<u>313</u>
7.5.4 Spline Estimate of the Density Approximation by Splines of the Even Order	<u>314</u>
7.6 Nonclassical Estimators	<u>315</u>
7.6.1 Estimators for the Distribution Function	<u>315</u>
7.6.2 Polygon Approximation of Distribution Function	<u>316</u>
7.6.3 Kernel Density Estimator	<u>316</u>
7.6.4 Projection Method of the Density Estimator	<u>318</u>
7.7 Asymptotic Rate of Convergence for Smooth Density Functions	<u>319</u>
7.8 Proof of Theorem 7.4	<u>322</u>
7.9 Choosing a Value of Smoothing (Regularization) Parameter for the Problem of Density Estimation	<u>327</u>
7.10 Estimation of the Ratio of Two Densities	<u>330</u>
7.10.1 Estimation of Conditional Densities	<u>333</u>
7.10.2 Estimation of Ratio of Two Densities on the Line	<u>334</u>
7.10.3 Estimation of a Conditional Probability on a Line	<u>337</u>
8 Estimating the Values of Function at Given Points	<u>339</u>
8.1 The Scheme of Minimizing the Overall Risk	<u>339</u>
8.2 The Method of Structural Minimization of the Overall Risk	<u>343</u>
8.3 Bounds on the Uniform Relative Deviation of Frequencies in Two Subsamples	<u>344</u>
8.4 A Bound on the Uniform Relative Deviation of Means in Two Subsamples	<u>347</u>

8.5 Estimation of Values of an Indicator Function in a Class of Linear Decision Rules	<u>350</u>
8.6 Sample Selection for Estimating the Values of an Indicator Function	<u>355</u>
8.7 Estimation of Values of a Real Function in the Class of Functions Linear in their Parameters	<u>359</u>
8.8 Sample Selection for Estimation of Values of Real-Valued Functions	<u>362</u>
8.9 Local Algorithms for Estimating Values of an Indicator Function	<u>363</u>
8.10 Local Algorithms for Estimating Values of a Real-Valued Function	<u>365</u>
8.11 The Problem of Finding the Best Point in a Given Set	<u>367</u>
8.11.1 Choice of the Most Probable Representative of the First Class	<u>368</u>
8.11.2 Choice of the Best Point of a Given Set	<u>370</u>
II Support Vector Estimation of Functions	
9 Perceptrons and Their Generalizations	<u>375</u>
9.1 Rosenblatt's Perceptron	<u>375</u>
9.2 Proofs of the Theorems	<u>380</u>
9.2.1 Proof of Novikoff Theorem	<u>380</u>
9.2.2 Proof of Theorem 9.3	<u>382</u>
9.3 Method of Stochastic Approximation and Sigmoid Approximation of Indicator Functions	<u>383</u>
9.3.1 Method of Stochastic Approximation	<u>384</u>
9.3.2 Sigmoid Approximations of Indicator Functions	<u>385</u>
9.4 Method of Potential Functions and Radial Basis Functions	<u>387</u>
9.4.1 Method of Potential Functions in Asymptotic Learning Theory	<u>388</u>
9.4.2 Radial Basis Function Method	<u>389</u>
9.5 Three Theorems of Optimization Theory	<u>390</u>
9.5.1 Fermat's Theorem (1629)	<u>390</u>
9.5.2 Lagrange Multipliers Rule (1788)	<u>391</u>
9.5.3 Kuhn-Tucker Theorem (1951) /	<u>393</u>
9.6 Neural Networks	<u>395</u>
9.6.1 The Back-Propagation Method	<u>395</u>
9.6.2 The Back-Propagation Algorithm	<u>398</u>

9.6.3 Neural Networks for the Regression Estimation Problem	<u>399</u>
9.6.4 Remarks on the Back-Propagation Method	<u>399</u>
10 The Support Vector Method for Estimating Indicator Functions	<u>401</u>
10.1 The Optimal Hyperplane	<u>401</u>
10.2 The Optimal Hyperplane for Nonseparable Sets	<u>408</u>
10.2.1 The Hard Margin Generalization of the Optimal Hyperplane	<u>408</u>
10.2.2 The Basic Solution Soft Margin Generalization	<u>411</u>
10.3 Statistical Properties of the Optimal Hyperplane	<u>412</u>
10.4 Proofs of the Theorems	<u>415</u>
10.4.1 Proof of Theorem 103	<u>415</u>
10.4.2 Proof of Theorem 104	<u>415</u>
10.4.3 Leave-One-Out Procedure	<u>416</u>
10.4.4 Proof of Theorem 105 and Theorem 9.2	<u>417</u>
10.4.5 Proof of Theorem 106	<u>418</u>
10.4.6 Proof of Theorem 107	<u>421</u>
10.5 The Idea of the Support Vector Machine	<u>421</u>
10.5.1 Generalization in High-Dimensional Space	<u>422</u>
10.5.2 Hilbert-Schmidt Theory and Mercer Theorem	<u>423</u>
10.5.3 Constructing SV Machines	<u>424</u>
10.6 One More Approach to the Support Vector Method	<u>426</u>
10.6.1 Minimizing the Number of Support Vectors	<u>426</u>
10.6.2 Generalization for the Nonseparable Case	<u>427</u>
10.6.3 Linear Optimization Method for SV Machines	<u>427</u>
10.7 Selection of SV Machine Using Bounds	<u>428</u>
10.8 Examples of SV Machines for Pattern Recognition	<u>430</u>
10.8.1 Polynomial Support Vector Machines	<u>430</u>
10.8.2 Radial Basis Function SV Machines	<u>431</u>
10.8.3 Two-Layer Neural SV Machines	<u>432</u>
10.9 Support Vector Method for Transductive Inference	<u>434</u>
10.10 Multiclass Classification	<u>437</u>
10.11 Remarks on Generalization of the SV Method	<u>440</u>
11 The Support Vector Method for Estimating Real-Valued Functions	<u>443</u>
11.1 ϵ -Insensitive Loss Functions	<u>443</u>
11.2 Loss Functions for Robust Estimators	<u>445</u>

11.3 Minimizing the Risk with ϵ -Insensitive Loss Functions	<u>448</u>
11.3.1 Minimizing the Risk for a Fixed Element of the Structure	<u>449</u>
11.3.2 The Basic Solutions	<u>452</u>
11.3.3 Solution for the Huber Loss Function	<u>453</u>
11.4 SV Machines for Function Estimation	<u>454</u>
11.4.1 Minimizing the Risk For a Fixed Element of the Structure in Feature Space	<u>455</u>
11.4.2 The Basic Solutions in Feature Space	<u>456</u>
11.4.3 Solution for Huber Loss Function in Feature Space	<u>458</u>
11.4.4 Linear Optimization Method	<u>459</u>
11.4.5 Multi-Kernel Decomposition of Functions	<u>459</u>
11.5 Constructing Kernels for Estimation of Real-Valued Functions	<u>460</u>
11.5.1 Kernels Generating Expansion on Polynomials	<u>461</u>
11.5.2 Constructing Multidimensional Kernels	<u>462</u>
11.6 Kernels Generating Splines	<u>464</u>
11.6.1 Spline of Order d with a Finite Number of Knots	<u>464</u>
11.6.2 Kernels Generating Splines with an Infinite Number of Knots	<u>465</u>
11.6.3 B_d -Spline Approximations	<u>466</u>
11.6.4 B_d -Splines with an Infinite Number of Knots	<u>468</u>
11.7 Kernels Generating Fourier Expansions	<u>468</u>
11.7.1 Kernels for Regularized Fourier Expansions	<u>469</u>
11.8 The Support Vector ANOVA Decomposition (SVAD) for Function Approximation and Regression Estimation	<u>471</u>
11.9 SV Method for Solving Linear Operator Equations	<u>473</u>
11.9.1 The SV Method	<u>473</u>
11.9.2 Regularization by Choosing Parameters of ϵ -Insensitivity	<u>478</u>
11.10 SV Method of Density Estimation	<u>479</u>
11.10.1 Spline Approximation of a Density	<u>480</u>
11.10.2 Approximation of a Density with Gaussian Mixture	<u>481</u>
11.11 Estimation of Conditional Probability and Conditional Density Functions	<u>484</u>
11.11.1 Estimation of Conditional Probability Functions	<u>484</u>
11.11.2 Estimation of Conditional Density Functions	<u>488</u>
11.12 Connections Between the SV Method and Sparse Function Approximation	<u>489</u>
11.12.1 Reproducing Kernels Hilbert Spaces	<u>490</u>
11.12.2 Modified Sparse Approximation and its Relation to SV Machines	<u>491</u>

12 SV Machines for Pattern Recognition	<u>493</u>
12.1 The Quadratic Optimization Problem	<u>493</u>
12.1.1 Iterative Procedure for Specifying Support Vectors	<u>494</u>
12.1.2 Methods for Solving the Reduced Optimization Problem	<u>496</u>
12.2 Digit Recognition Problem: The U.S. Postal Service Database	<u>496</u>
12.2.1 Performance for the U.S. Postal Service Database	<u>496</u>
12.2.2 Some Important Details	<u>500</u>
12.2.3 Comparison of Performance of the SV Machine with Gaussian Kernel to the Gaussian RBF Network	<u>503</u>
12.2.4 The Best Results for U.S. Postal Service Database	<u>505</u>
12.3 Tangent Distance	<u>506</u>
12.4 Digit Recognition Problem: The NIST Database	<u>511</u>
12.4.1 Performance for NIST Database	<u>511</u>
12.4.2 Further Improvement	<u>512</u>
12.4.3 The Best Results for NIST Database	<u>512</u>
12.5 Future Racing	<u>514</u>
12.5.1 One More Opportunity: The Transductive Inference	<u>518</u>
13 SV Machines for Function Approximations, Regression Estimation, and Signal Processing	<u>521</u>
13.1 The Model Selection Problem	<u>521</u>
13.1.1 Functional for Model Selection Based on the VC Bound	<u>522</u>
13.1.2 Classical Functionals	<u>524</u>
13.1.3 Experimental Comparison of Model Selection Methods	<u>525</u>
13.1.4 The Problem of Feature Selection Has No General Solution	<u>526</u>
13.2 Structure on the Set of Regularized Linear Functions	<u>530</u>
13.2.1 The <i>L</i> -Curve Method	<u>532</u>
13.2.2 The Method of Effective Number of Parameters	<u>534</u>
13.2.3 The Method of Effective VC Dimension	<u>536</u>
13.2.4 Experiments on Measuring the Effective VC Dimension	<u>540</u>
13.3 Function Approximation Using the SV Method	<u>543</u>

13.3.1 Why Does the Value of ϵ Control the Number of Support Vectors?	<u>546</u>
13.4 SV Machine for Regression Estimation	<u>549</u>
13.4.1 Problem of Data Smoothing	<u>549</u>
13.4.2 Estimation of Linear Regression Functions	<u>550</u>
13.4.3 Estimation of Nonlinear Regression Functions	<u>556</u>
13.5 SV Method for Solving the Positron Emission Tomography (PET) Problem	<u>558</u>
13.5.1 Description of PET	<u>558</u>
13.5.2 Problem of Solving the Radon Equation	<u>560</u>
13.5.3 Generalization of the Residual Principle of Solving PET Problems	<u>561</u>
13.5.4 The Classical Methods of Solving the PET Problem	<u>562</u>
13.5.5 The SV Method for Solving the PET Problem	<u>563</u>
13.6 Remark About the SV Method	<u>567</u>
III Statistical Foundation of Learning Theory	
14 Necessary and Sufficient Conditions for Uniform Convergence of Frequencies to their Probabilities	<u>571</u>
14.1 Uniform Convergence of Frequencies to their Probabilities	<u>572</u>
14.2 Basic Lemma	<u>573</u>
14.3 Entropy of the Set of Events	<u>576</u>
14.4 Asymptotic Properties of the Entropy	<u>578</u>
14.5 Necessary and Sufficient Conditions of Uniform Convergence Pmof of Sufficiency	<u>584</u>
14.6 Necessary and Sufficient Conditions of Uniform Convergence Pmof of Necessity	<u>587</u>
14.7 Necessary and Sufficient Conditions Continuation of Proving Necessity	<u>592</u>
15 Necessary and Sufficient Conditions for Uniform Convergence of Means to their Expectations	<u>597</u>
15.1 ϵ -Entropy	<u>597</u>
15.1.1 Root of the Existence of the Limit	<u>600</u>
15.1.2 Root of the Convergence of the Sequence	<u>601</u>
15.2 The Quasicube	<u>603</u>
15.3 ϵ -Extension of a Set	<u>608</u>

15.4 An Auxiliary Lemma	<u>610</u>
15.5 Necessary and Sufficient Conditions for Uniform Convergence: The Proof of Necessity	<u>614</u>
15.6 Necessary and Sufficient Conditions for Uniform Convergence: The Proof of Sufficiency	<u>618</u>
15.7 Corollaries from Theorem 15.1	<u>624</u>
16 Necessary and Sufficient Conditions for Uniform One-Sided Convergence of Means to Their Expectations	<u>629</u>
16.1 Introduction	<u>629</u>
16.2 Maximum Volume Sections	<u>630</u>
16.3 The Theorem on the Average Logarithm	<u>636</u>
16.4 Theorem on the Existence of a Corridor	<u>642</u>
16.5 Theorem on the Existence of Functions Close to the Corridor Boundaries (Theorem on Potential Nonfalsifiability)	<u>651</u>
16.6 The Necessary Conditions	<u>660</u>
16.7 The Necessary and Sufficient Conditions	<u>666</u>
Comments and Bibliographical Remarks	<u>681</u>
References	<u>723</u>
Index	<u>733</u>

PREFACE

This book is devoted to statistical learning theory, the theory that explores ways of estimating functional dependency from a given collection of data. This problem is very general. It covers important topics of classical statistics—in particular, discriminant analysis, regression analysis, and the density estimation problem.

In this book we consider a new paradigm for solving these problems: the so-called learning paradigm that was developed over the last 30 years. In contrast to the classical statistics developed for large samples and based on using various types of a priori information, the new theory was developed for small data samples and does not rely on a priori knowledge about a problem to be solved. Instead it considers a structure on the set of functions implemented by the learning machine (a set of nested subsets of functions) where a specific measure of subset capacity is defined.

To control the generalization in the framework of this paradigm, one has to take into account two factors, namely, the quality of approximation of given data by the chosen function and the capacity of the subset of functions from which the approximating function was chosen.

This book presents a comprehensive study of this type of inference (learning process). It contains:

- The general qualitative theory that includes the necessary and sufficient conditions for consistency of learning processes
- The general quantitative theory that includes bounds on the rate of convergence (the rate of generalization) of these learning processes
- Principles for estimating functions from a small collection of data that are based on the developed theory
- Methods of function estimation and their application to solving real-life problems that are based on these principles

The book has three parts: "Theory of Learning and Generalization," "Support Vector Estimation of Functions," and "Statistical Foundation of Learning Theory."

The first part, "Theory of Learning and Generalization," analyzes factors

responsible for generalization and shows how to control these factors in order to generalize well.

This part contains eight chapters. Chapter 1 describes two different approaches to the learning problem. The first approach considers learning as a problem of minimizing an expected risk functional in the situation when the probability measure that defines the risk is unknown but i.i.d. observations are given. To obtain a solution in the framework of this approach, one has to suggest some inductive principle. That is, one has to define a constructive functional that should be minimized (instead of the expected risk functional) in order to find a function that guarantees a small expected loss. The second approach considers learning as a problem of identification of the desired function: Using observations, one has to find the function that is close to the desired one. In general, this approach leads to the necessity of solving the so-called ill-posed problems.

Chapter 2 discusses connections between the main problems of learning theory and problems of the foundation of statistics, namely the problem of estimating the probability measure from the data. It describes two ways of estimating the probability measure. One way is based on the convergence of an estimate of the probability measure in a weak mode, and another way is based on convergence in a strong mode. These two ways of estimating the unknown measure imply two approaches to the learning problem described in Chapter 1.

Chapter 3 is devoted to the qualitative model of learning processes, namely, to the theory of consistency of the learning processes based on the empirical risk minimization induction principle. It shows that for consistency of the learning processes based on this principle the convergence of some empirical processes (the existence of uniform law of large numbers) is necessary and sufficient. In Chapter 3 these conditions are discussed. (The corresponding theorems will be proven in the third part of the book.)

Chapters 4 and 5 estimate the bounds on the rate of convergence of the empirical processes. Using these bounds we obtain bounds on the risk for the functions that minimize the empirical risk functional. In Chapter 4 we obtain bounds for sets of indicator functions (for the pattern recognition problem), and in Chapter 5 we generalize these bounds for sets of real-valued functions (for regression estimation problems). The bounds depend on two factors: the value of empirical risk and the capacity of the set of functions from which the function minimizing empirical risk was chosen.

In Chapter 6 we introduce a new induction principle, the so-called "structural risk minimization" principle, which minimizes bounds obtained in Chapters 4 and 5 with respect to two factors, the value of empirical risk and the capacity. This principle allows us to find the function that achieves the guaranteed minimum of the expected risk using a finite number of observations.

Chapter 7 is devoted to solving stochastic ill-posed problems, including the problems of density estimation, conditional density estimation, and conditional probability estimation. For solving these problems, we utilize the

regularization method (which is based on the same ideas as the structural risk minimization principle). Using this method, we obtain both the classical methods for the solution of our problems and new ones.

In Chapter 8 we consider a new statement of the learning problem. We introduce the problem of estimating values of a function at given points of interest. For a restricted amount of empirical data, the generalization ability using the direct methods of estimating the values of a function at given points of interest can be better than using methods of estimating the function. Therefore, we consider methods of direct estimation of the values of the function at given points of interest that are not based on the estimation of the functional dependency.

The second part of this book, "Support Vector Estimation of Functions," introduces methods that provide generalization when estimating a multi-dimensional function from a limited collection of data.

This part contains five chapters. Chapter 9 describes classical algorithms: Perceptrons, neural networks, and radial basis functions.

Chapters 10, 11, 12, and 13 are devoted to new methods of solving dependency estimation problems, the so-called support vector method. Chapter 10 considers support vector machines for estimating indicator functions (for pattern recognition problems). Chapter 11 considers support vector machines for estimating real-valued functions.

Chapters 12 and 13 discuss solutions of real-life problems using support vector machines. Chapter 12 discusses pattern recognition problems, and Chapter 13 discusses various real-valued function estimation problems such as function approximation, regression estimation, and solving inverse problems.

The third part of this book "Statistical Foundation of Learning Theory," studies uniform laws of large numbers that make generalization possible.

This part contains three chapters. Each of these chapters studies a different empirical process: uniform convergence of frequencies to their probabilities over a given set of events (Chapter 14), uniform convergence of means to their expectations over a given set of functions (Chapter 15), and uniform one-sided convergence of means to their expectations over a given set of functions (Chapter 16). Convergence of these processes forms the basis for the theory of learning processes and for theoretical statistics.

Bibliographical, historical, and general comments, reflecting the author's point of view on the development of statistical learning theory and related disciplines, are given at the end of the book.

The first two parts of the book are written at a level for use in a graduate course on learning theory in statistics, mathematics, engineering, physics, and computer science. It should also appeal to professional engineers wishing to learn about learning theory or to use new methods for solving real-life problems. The third part is written at a higher level. It can be used in a special course on empirical processes for Ph.D. students in mathematics and statistics.

This book became possible due to the support of Larry Jackel, the head of the Adaptive System Research Department. AT&T Bell Laboratories and Yann LeCun, the head of the Image Processing Research Department. AT&T Research Laboratories. It was inspired by collaboration with my colleagues Yoshua Bengio, Bernhard Boser, Leon Bottou. Chris Burges, Eric Cosatto. John Denker. Harris Drucker. Alexander Gammerman, Hans Peter Craft, Isabella Guyon, Patrick Haffner, Martin Hasler. Larry Jackel, Yann LeCun, Esther Levin, Robert Lyons. Nada Matic, Craig Nohl, Edwin Pednault, Edward Sackinger. Bernard Schiilkopf. Alex Smola, Patrice Simard. Sara Solla. Vladimir Vovk, and Chris Watkins.

I discussed the ideas presented in the book with Leo Breiman. Jerry Friedman, Federico Girosi, Tomaso Poggio, Yakov Kogan, and Alexander Shustorovich. The feedback of these discussions had an important influence on the content of this book.

Martin Hasler, Federico Girosi, Edward Rietman. Pavel Laskov, Mark Stitson, and Jason Weston read the manuscript and improved and simplified the presentation.

When the manuscript was completed I gave some chapters to Chris Burges, Leon Bottou. Hamid Jafarkhani, Ilia Izmailov, Art Owen, Bernhard Schölkopf, and Michael Turmon. They also improved the quality of the book.

I would like to express my deep gratitude to all of them.

VLADIMIR N. VAPNIK

*Red Bank, New Jersey,
June 1998*

INTRODUCTION: THE PROBLEM OF INDUCTION AND STATISTICAL INFERENCE

0.1 LEARNING PARADIGM IN STATISTICS

The goal of this book is to describe a new approach to dependency estimation problems which originated within learning theory.

The development of this approach started in the 1960s after the appearance of the first generation of computers capable of conducting multidimensional analysis of real-life problems. From the very first results of these analyses it became clear that existing classical approaches to low-dimensional function estimation problems do not reflect singularities of high-dimensional cases. There was something in high-dimensional cases that was not captured by the classical paradigm. R. Bellman called this something "the curse of dimensionality." In attempts to overcome this curse a new paradigm was developed.

When developing the new paradigm it was fortunate that in the late 1950s F. Rosenblatt started analysis of the pattern recognition problem. From the formal point of view the pattern recognition problem belongs to the general statistical problem of function estimation from empirical data. However, in this problem one has to estimate a function belonging to simple sets of functions—sets of indicator functions. Analysis of these simple sets was crucial for discovery of the concepts that determine the generalization ability, the so-called capacity concepts of a set of functions. These concepts would be hard to extract from analysis of more sophisticated sets of functions—sets of real-valued functions. Capacity control became one of the main tools in the new approach.

Later, in the 1980s, when the theory of this approach had been essentially developed, it was noted that a generalized version of one of the problems at the cornerstone of statistics (the Glivenko–Cantelli problem) leads to the same analysis that was developed for the theory of learning and generaliza-

tion in pattern recognition. In the mid-1980s these results were rewritten in traditional statistical terms. Nevertheless, the new paradigm in statistics was developed at the periphery of statistical science as an attempt to analyze the problem of generalization in the simplest model of statistical inference—the pattern recognition problem.

This fact constitutes an important methodological discovery. The pattern recognition problem is one of the simplest models of inductive inference. Results for this model can be generalized for other (more complex) models using more or less standard mathematical techniques. Therefore in studies of statistical inference, the pattern recognition model plays the same role as the drosophila fly in studies of genetic structures.

In this book we try to develop a general approach to statistical inference. For this purpose we analyze the pattern recognition problem in great detail and then generalize the obtained results for solving main problems of statistical inference.

0.2 TWO APPROACHES TO STATISTICAL INFERENCE: PARTICULAR (PARAMETRIC INFERENCE) AND GENERAL (NONPARAMETRIC INFERENCE)

The elements of statistical inference have existed for more than 200 years (one can find them in the works of Gauss and Laplace): however, the systematic analysis of these problems started only in the late 1920s.

By that time, descriptive statistics was mostly complete: It was shown that there are different statistical laws (distribution functions) that describe well many events of reality. The next question to be investigated was finding a reliable method of statistical inference. The problem was as follows:

Given a collection of empirical data originating from some functional dependency, infer this dependency.

In the 1920s the analysis of methods of statistical inference began. Two bright events signaled this start:

1. Fisher introduced the main models of statistical inference in the unified framework of parametric statistics. He described different problems of estimating functions from given data (the problems of discriminant analysis, regression analysis, and density estimation) as the problems of parameter estimation of specific (parametric) models and suggested one method for estimating the unknown parameters in all these models—the maximum likelihood method.
2. Glivenko, Cantelli, and Kolmogorov started a general analysis of statistical inference. Glivenko and Cantelli proved that the empirical distribution function always converges to the actual distribution function. Kolmogorov found the asymptotically exact rate of this convergence.

The rate turns out to be fast (exponential) and independent of the unknown distribution function.

These two events determined two main approaches to statistical inference:

1. The *particular* (parametric) inference, which aims to create simple statistical methods of inference that can be used for solving real-life problems, and
2. The *general* inference, which aims to find one (induction) method for any problem of statistical inference.

The philosophy that led to the creation of parametric statistical inference is based on the following belief:

The investigator knows the problem to be analyzed rather well. He knows the physical law that generates the stochastic properties of the data and the function to be found up to a finite number of parameters. Estimating these parameters using the data is considered to be the essence of the problem of statistical inference. To find these parameters using information about the statistical law and the target function, one adopts the maximum likelihood method.

The goal of the theory is to justify this approach (by discovering and describing its favorable properties).

The philosophy that led to general statistical inference is different:

One does not have reliable a priori information about the statistical law underlying the problem or about the function that one would like to approximate. It is necessary to find a method to infer an approximating function from the given examples in this situation.

The corresponding theory must:

1. Describe conditions under which one can find in a given set of functions the best approximation to an unknown function with an increasing number of examples.
2. Find the best method of inference for a given number of examples.

Kolmogorov's discovery that the empirical distribution function has a universally (i.e., independent of the actual distribution function) asymptotic exponential rate of convergence fostered hope that the general type of inference is feasible. The results of Glivenko, Cantelli, and Kolmogorov started more than 40 years of research on general statistical inference before it culminated in inductive methods.

The theory of these methods is the subject of this book.

0.3 THE PARADIGM CREATED BY THE PARAMETRIC APPROACH

In contrast to the slow development of general inductive inference, the parametric approach to inductive inference was developed very quickly. In fact, the main ideas of parametric inference were developed in the 1930s, and during the next 10 years the main elements of the theory of parametric inference were introduced.

The 30-year period between 1930 and 1960 can be called the "golden age" of parametric inference. During this period, one approach to statistical inference dominated: the approach based on parametric paradigms. Only one theory of statistical inference was accepted, namely the theory that served the parametric paradigm.

Of course, the results of Glivenko, Cantelli, and Kolmogorov were known; however, they were considered as inner technical achievements that are necessary for the foundation of statistical theory rather than an indication that there could be a different type of inference which is more general and more powerful than parametric inference.

In any case, almost all standard statistical textbooks considered the problem of inference from the point of view of the parametric paradigm, and thus several generations of statisticians were educated in this framework.[†]

The philosophy of the classical parametric paradigm is based on the following three beliefs:

1. *To find a functional dependency from the data, the statistician is able to define a set of functions, linear in their parameters, that contain a good approximation to the desired function. The number of free parameters describing this set is small.*

This belief was supported by referring to the Weierstrass theorem, according to which any continuous function can be approximated on a finite interval by polynomials (functions linear in their parameters) with any degree of accuracy. The idea was that if polynomials can approximate the desired function well, then a smart statistician can define a set of functions, linear in their parameters (not necessarily polynomials) with a small number of free parameters that provides a good approximation to the desired function.

2. *The statistical law underlying the stochastic component of most real-life problems is the normal law.*

This belief was supported by referring to the Central Limit Theorem, which states that under wide conditions the sum of a large number of random

[†] It is fair to note that in the time before wide availability of computers (before 1960s) the goal of applied statistics was to create computationally simple methods, and parametric statistics was responsive to these limitations.

variables is approximated by the normal law. The idea was that if randomness in the problem is the result of interaction among a large number of random components, then the stochastic element of the problem is described by the normal law.

3. *The induction engine in this paradigm—the maximum likelihood method—is a good tool for estimating parameters.*

This belief was supported by many theorems about *conditional optimality* of the method (optimality in a restricted set of methods and/or in the asymptotic case). The maximum likelihood method was hoped to be a good tool for estimating parameters of models even for small sample sizes.

Note that these three beliefs were also supported by the philosophy:

If there exists a mathematical proof that some method provides an asymptotically optimal solution, then in real life this method will provide a reasonable solution for a small number of data samples.

0.4 SHORTCOMING OF THE PARAMETRIC PARADIGM

In the 1960s, the wide application of computers for solving scientific and applied problems started. Using computers, researchers for the first time tried to analyze sophisticated models (that had many factors) or tried to obtain more precise approximations. These efforts immediately revealed shortcomings of the parametric paradigm in all three of the beliefs upon which the paradigm was based.

1. First, the computer analysis of large multivariate problems resulted in the discovery of the phenomenon that R. Bellman called "the curse of dimensionality." It was observed that increasing the number of factors that have to be taken into consideration requires exponentially increasing the amount of computational resources. For example, according to the Weierstrass theorem, any continuous function (of n variables) defined on the unit cube can be approximated by polynomials with any degree of accuracy. However, if the desired function has only s derivatives, then using polynomials with N terms one can only guarantee the accuracy $O(N^{-s/n})$. If the unknown function is not very smooth (i.e., it possesses only a small number of derivatives), then to obtain the desired level of accuracy one needs an exponentially increasing number of terms with an increasing number, n , of variables.

Therefore, in real-life multidimensional problems in which one may consider dozens or even hundreds of variables, the belief that one can define a reasonably small set of functions that contains a good approximation to a desired one looks naive.

2. Approximately at the same time, by analyzing real-life data, Tukey demonstrated that the statistical components of real-life problems cannot be described by only classical statistical distribution functions. Often real-life distributions are different, and one must take this difference into account in order to construct effective algorithms.
3. In addition, James and Stein showed that even for simple problems of density estimation, such as the problem of estimating the location parameters of $n > 2$ dimensional normal law with unit covariance matrix (for estimating means), the maximum likelihood method is not the best one. They suggested an estimator that for this specific problem is uniformly better than the maximum likelihood estimator.

Thus, all three beliefs on which the classical paradigm relied turned out to be inappropriate for many real-life problems. This had an enormous consequence for statistical science: It looked as if the idea of constructing statistical inductive inference methods for real-life problems had failed.

0.5 AFTER THE CLASSICAL PARADIGM

The discovery of difficulties with the classical paradigm was a turning point in statistics. Many statisticians reconsidered the main goal of the entire statistical analysis business. A new direction in statistics was declared, the so-called "data analysis," where the goal was to help researchers perform inductive inferences from data, rather than to do so using purely statistical techniques. Therefore, various techniques were developed for visualizing data, for clustering data, for constructing features, and so on. In other words, tools were developed that would enable a researcher to make *informal* inferences.

One can summarize the philosophy of the data analysis approach as the following declaration:

Inductive inference is an informal act, and statisticians contribute to this act only by technical assistance.

One must note, however, that tremendous efforts have been made to save the classical paradigm by generalizing all three of its main presumptions:

1. In the 1960s, P. Huber developed the so-called *robust* approach to parametric statistics, where one does not need to specify a statistical law in order to estimate a function from a given parametric set of functions.
2. In the 1970s, in an attempt to use a wider set of functions, J. Nelder and R. Wedderburn suggested the so-called *generalized linear models*. Attempts to use wide sets of functions created the problem of *model selection*. Several asymptotic results regarding solutions of this problem

were obtained. (However, understanding of the model-selection problem as a small sample size problem came later when a new inductive paradigm was created. We will discuss the small sample size problem in this book.)

3. In the 1980s L. Breiman, J. Huber, and J. Friedman started to consider special types of functions, nonlinear in their parameters, and started to use the regularized empirical risk minimization method instead of the maximum likelihood method.

Nevertheless, in spite of these and many other achievements, the limitations of the classical parametric paradigm remain, and therefore currently not many researchers consider the classical paradigm as the main approach to statistical inference.

0.6 THE RENAISSANCE

The return to the general problem of statistical inference occurred so imperceptibly that it was not recognized for more than 20 years.

In 1958, F. Rosenblatt, a physiologist, suggested a learning machine (computer program) called the Perceptron for solving the simplest learning problem: namely, the classification (pattern recognition) problem. The construction of this machine reflected some existing neurophysiological models of learning mechanisms. With the simplest examples, F. Rosenblatt demonstrated that the Perceptron could generalize. After the Perceptron, many different types of learning machines were suggested. They didn't generalize worse than the Perceptron, but they had no neurobiological analogy.

The natural question arose:

Does there exist something common in these machines? Does there exist a general principle of inductive inference that they implement?

Immediately a candidate was found for such a general induction principle: the so-called empirical risk minimization (ERM) principle. In order to achieve good generalization on future (test) examples, the ERM principle suggests a decision rule (an indicator function) that minimizes the number of training errors (empirical risk). The problem was to construct a theory for this principle.

At the end of the 1960s, the theory of ERM for the pattern recognition problem was constructed.[†] This theory included both (a) the general *qualitative theory* of generalization that described the necessary and sufficient

[†]See monograph by V. N. Vapnik and A. Ya. Chervonenkis *Theory of Pattern Recognition*. Nauka, Moscow, 1974, 416 pages. German translation: W. N. Vapnik and A. Ya. Tscherwonenskis *Theorie der Zeichenerkennung*, Akademia, Berlin, 1979, 352 pages.

conditions for consistency of the ERM induction principle (valid for any set of indicator functions—that is, $\{0, 1\}$ valued functions on which the machine minimizes the empirical risk) and (b) the general *quantitative theory* that described the bounds on the probability of the (future) test error for the function minimizing the empirical risk.

It must be noted that the ERM principle was discussed in the statistical literature several times before. The essential difference, however, was that in the pattern recognition problem ERM inference is applied to sets of *simple functions*—namely to sets of *indicator* functions—while in classical statistics it was applied to various sets of real-valued functions. Within 10 years, the theory of the ERM principle was generalized for sets of real-valued functions as well.¹ However, it was extremely lucky that at the first and the most important stage of developing the theory, when the main concepts of the entire theory had to be defined, simple sets of functions were considered. Generalizing the results obtained for estimating indicator functions (pattern recognition) to the problem of estimating real-valued functions (regressions, density functions, etc.) was a purely technical achievement. To obtain these generalizations, no additional concepts needed to be introduced.

0.7 THE GENERALIZATION OF THE GLIVENKO-CANTELLI-KOLMOGOROV THEORY

Application of the ERM principle does not necessarily guarantee consistency (i.e., convergence to the best possible solution with an increasing number of observations). Therefore, the main issues that drove the development of the ERM theory were as follows:

1. To describe situations under which the method is consistent—that is, to find the necessary and sufficient conditions for which the ERM method defines functions that converge to the best possible solution with an increasing number of observations. The resulting theorems thereby describe the qualitative model of ERM inference.
2. To estimate the quality of the solution obtained on the basis of the given sample size—that is, to estimate both the probability of error for the function that minimizes the empirical risk on the given set of training examples and to estimate how close this probability is to the smallest possible for the given set of functions. The resulting theorems characterize the generalization ability of the ERM principle.

To address both these issues for the pattern recognition problem, it was

¹ See monograph by V. N. Vapnik *Estimation of Dependencies Based on Empirical Data*, Nauka, Moscow, 1979, 442 pages. English translation: Springer-Verlag, New York, 1982, 500 pages.

necessary to construct a theory that can be considered as a generalization of the Glivenko–Cantelli–Kolmogorov results.

According to the classical law of large numbers, the frequency of any event converges to the probability of this event with an increasing number of observations. However, the classical law of large numbers is not sufficient to assert that for a given set of events the sequence of probabilities of events with the *smallest frequency* converges to the *smallest possible value for this set* (i.e., to assert the consistency of the ERM method). Instead, it was proven that in order to ensure the consistency of the ERM method, it is *necessary and sufficient* that the *uniform law of large numbers* holds (uniform over all events of the set of events defined by the set of indicator functions implemented by the learning machine).

One can reformulate the Glivenko–Cantelli theorem as an assertion that for some *specific set of events* there exists a uniform law of large numbers and the Kolmogorov's bound as the bound on the asymptotic rate of *uniform convergence* of the frequencies to their probabilities over this specific set of events. Therefore, to construct a general theory of the ERM method for pattern recognition, one has to generalize the Glivenko–Cantelli–Kolmogorov theory; that is:

1. For any given set of events, to determine whether the uniform law of large numbers holds (i.e., does uniform convergence take place!).
2. If uniform convergence holds, to find the bounds for the *nonasymptotic* rate of uniform convergence.

Note that these bounds are generalizations of Kolmogorov's bound in two respects: They must be valid for a finite number of observations and they must be valid for any set of events.

This theory was constructed in the late 1960s (Vapnik and Chervonenkis, 1968, 1971). The cornerstone in this theory is a collection of new concepts, the so-called capacity concepts for a set of events (a set of indicator functions). Of particular importance is the so-called VC dimension of the set of events (the VC dimension of the set of indicator functions implemented by the learning machine) which characterizes the variability of the set of events (indicator functions). It was found that both the necessary and sufficient conditions of consistency and the rate of convergence of the ERM principle depend on the capacity of the set of functions implemented by the learning machine.

In particular, it was proven that for distribution-independent consistency of the ERM principle, it is necessary and sufficient that the set of functions implemented by the learning machine has a finite VC dimension. It was also found that distribution-free bounds on the rate of uniform convergence depend on the VC dimension, the number of training errors, and the number of observations.

0.8 THE STRUCTURAL RISK MINIMIZATION PRINCIPLE

The bounds for the rate of uniform convergence not only provide the main theoretical basis for the ERM inference, but also motivate a new method of inductive inference.

For any level of confidence, an equivalent form of the bounds define bounds on the probability of the test error *simultaneously for all functions of the learning machine* as a function of the number of training errors, of the VC dimension of the set of functions implemented by the learning machine, and of the number of observations.

This form of the bounds led to a new idea for controlling the generalization ability of learning machines:

To achieve the smallest bound on the test error by controlling (minimizing) the number of training errors, the machine (the set of functions) with the smallest VC dimension should be used.

These two requirements—to minimize the number of training errors and to use a machine (a set of functions) with a small VC dimension—are contradictory: To minimize the number of training errors, one needs to choose a function from a wide set of functions, rather than from a narrow set, with small VC dimension. Therefore, to find the best guaranteed solution, one has to make a compromise between the accuracy of approximation of the training data and the capacity (the VC dimension) of the machine that one uses to minimize the number of errors. The idea of minimizing the test error by controlling two contradictory factors was formalized by introducing a new induction principle, the so-called Structural Risk Minimization (SRM) principle.¹

One has to note that the idea of the existence of a compromise in inductive inference has been discussed in philosophy for almost 700 years, since William of Occam proposed in the fourteenth century the general principle known as Occam's razor:

Entities should not be multiplied beyond necessity.

The attempt to provide Occam's razor with an exact sense underlies these discussions. The most common interpretation of Occam's razor is:

The simplest explanation is the best.

The assertion that comes from the SRM theory is different:

See footnote on page 8.

The explanation by the machine with the smallest capacity (VC dimension) is the best.

Two important points should be mentioned in connection with introducing the capacity concept (instead of the simplicity concept).

First, capacity determines both the *necessary and sufficient* conditions for consistency of learning processes and the rate of convergence of learning processes. Therefore, it reflects intrinsic properties of inductive inference.

Second, naive notions of complexity (for example, the number of parameters) do not necessarily reflect capacity properly. In this book, we will describe an example of a simple set of functions that depends on only one parameter and that has infinite VC dimension, as well as a set of functions with a billion parameters that has low VC dimension. We will see that if the VC dimension of a set of functions is infinite (even if we consider a set of "simple" functions), then the so-called situation of nonfalsifiability (described by K. Popper in his analysis of philosophy of science) prevents generalization from taking place. On the other hand, we will also describe a learning machine, which uses a high-order of polynomials (say five) in a high-dimensional space (say 400) which has a good generalization ability due to capacity control.

The discovery that the generalization ability of the learning machine depends on the capacity of the set of functions implemented by the learning machine which differ from the number of free parameters is one of the most important achievements of the new theory.

Capacity control in inductive inference makes it possible to take into account the amount of training data. This was discovered in the mid-1970s for the pattern recognition problem; and by the beginning of 1980, all of the results obtained for sets of indicator functions were generalized for sets of real-valued functions (for the problem of regression estimation).

Capacity control in a structured set of functions became the main tool of the new paradigm. It is especially important when one tries to make an inference based on a *small sample sizes*.[†]

0.9 THE MAIN PRINCIPLE OF INFERENCE FROM A SMALL SAMPLE SIZE

The key idea for creating effective methods of inference from small sample sizes is that one performs inference in situations where one possesses a restricted amount of information. To take this fact into account, we formulate the following Main Principle:

[†] We consider the size ℓ of data to be small for estimating functions on the basis of the set of functions with VC dimension h if the ratio ℓ/h is small (say $\ell/h < 20$).

If you possess a restricted amount of information for solving some problem, try to solve the problem directly and never solve a more general problem as an intermediate step. It is possible that the available information is sufficient for a direct solution but is insufficient for solving a more general intermediate problem.

In spite of the obviousness of the Main Principle, it is not easy to follow it. At least the classical approach to statistical inference does not follow this principle. Indeed, in order to estimate decision rules, the classical approach suggests estimating densities as a first step (recall the classical parametric paradigm based on the maximum likelihood method). Note that estimating probability densities is a universal problem of statistics. Knowing the density, one can solve many different problems. For example, one can estimate the conditional density, which can be described as a ratio of two densities. Therefore, in general, density estimation is a hard (ill-posed) problem that requires a large number of observations to be solved well.

However, even if one needs to estimate the conditional density, one must try to find it *directly*, and not as a ratio of two estimated densities. Note that often conditional densities can be approximated by low-dimensional functions even if the densities are high-dimensional functions.

In an attempt to solve the function estimation problem directly, we derived bounds on the quality of any possible solution (bounds on the generalization ability) and introduced a method to control the generalization ability by minimizing these bounds. This brought us to the SRM inductive principle which explicitly incorporates capacity control.

Following the logic of the Main Principle a step further brings us to an idea of inference that goes beyond induction.

In many real-life problems, the goal is to find the values of an unknown function only at points of interest (i.e., on the test set). To solve this problem, the established paradigm uses a two-stage procedure: At the first (induction) stage we estimate the function from a given set of functions using an induction principle, while at the second (deduction) stage we use this function to evaluate the values of the unknown function at the points of interest. At the first stage of this two-stage scheme, we thus solve a problem that is more general than the one we need to solve. To estimate an unknown function means to estimate its values at all points in the domain of this function. Why solve a much more general problem—function estimation—if we only need to estimate the values of a function at a few (> 1) points of interest? In situations where we have a restricted amount of information, it is possible that we can estimate the values of the unknown function reasonably well at given points of interest but cannot estimate the values of the function well at *all* points of its domain.

The direct estimation of values of a function only at points of interest using a given set of functions forms a new type of inference which can be called *transductive inference*. In contrast to the inductive solution that derives results

in two steps, from particular to general (the inductive step) and then from general to particular (the deductive step), the transductive solution derives results in one step, directly from particular to particular (the transductive step).

Therefore the classical paradigm often contradicts the Main Principle. To avoid these contradictions a new approach was developed.

0.10 WHAT THIS BOOK IS ABOUT

This book is devoted to the theory of inductive inference, a model of which is statistical inference (inference for the simplest statistical models).

The main problem in inductive inference lies in philosophy, in finding the principles of inference[†] rather than in the mathematical analysis of the formulated principles. However, to find the principles of inference that reflect the phenomenon of human inference, one cannot utilize two thousand years of philosophical heritage. Recall that when in the beginning of the 1960s the problem of modeling learning processes on computers arose, the only inspiration for constructing learning machines was a physiological analogy (the Perceptron), but not general philosophical principles.

For this reason, it is important to analyze in great detail a simple mathematical problem of induction and try to discover the general principles of inference from this analysis. Such a simple mathematical problem is the pattern recognition problem.[‡]

The following three claims constitute the most important results of analyzing the pattern recognition problem and its generalization, the estimation of real-valued functions:

1. *The theory of induction is based on the uniform law of large numbers.*
2. *Effective methods of inference must include capacity control.*
3. *Along with inductive inference there exists transductive inference which in many cases may be preferable.*

Not all of these claims are justified equally well.

[†]From this point of view, the methodology of research of inductive inference is similar to the methodology of physical science: There exists some phenomenon of nature for which a model should be found. The mathematical analysis presented here is a tool that helps one to find this model. The result of any analysis should be confirmed by experiments.

[‡]The simplest induction problem is estimating the function from a set of constants—that is, functions that take on only one value. This was the case actually under consideration when the classical theory was developed. However, the structure of the set of constant functions is too simple, since any subset of constant functions has the same VC dimension, equal to one. Therefore, the simplest model of induction that requires capacity control is the pattern recognition problem.

1. The analysis of the uniform law of large numbers and its relation to the problem of induction inference is almost complete. It includes both the qualitative analysis of the model (the analysis of the necessary and sufficient conditions for consistency) and the quantitative analysis of the model (the theory of bounds). The largest part of this book (Chapters 3, 4, 5, 14, 15, and 16) is devoted to this analysis.
2. In spite of the fact that the capacity control principle (the SRM principle) was discovered in the middle of the 1970s, the development of this principle—which led to new types of algorithms, the so-called Support Vector Machines (SVM)—started only in the 1990s. So far, we have only the first results of the theoretical analysis, along with the first results of practical applications. Chapters 6, 10, 11, 12, and 13 are devoted to this subject. Chapter 7 is closely related to capacity control methods. It describes a theory of stochastic ill-posed problems and its application to the problem of density and conditional density estimation.
7. Lastly, the theory of transductive inference is only at a very early stage of development. We have described only very general combinatorial ideas on factorizing a given set of functions based on a given set of data. However, new methods for capacity control developed in the last few years (described in Chapters 10, 11, 12, and 13) appear to be a good tool for implementing transductive inference. Only one chapter (Chapter 8) is devoted to analysis of this type of inference.

In spite of the fact that this book explicitly deals only with the mathematical problems of inductive inference, it implicitly contains two additional subjects of discussion: (1) a discussion of the general problem of induction and (2) a discussion of the existence of various methods of inference, namely, inference through induction (generalization) and inference through transduction, the direct (ad hoc) inference. In Chapter 3, there is a direct comparison of the capacity concepts with some fundamental concepts developed by K. Popper in the philosophy of science. The problem of transductive inference has no such remarkable achievement in philosophy as Popper's theory. The existence of a direct type of inference is still under discussion in philosophy. Therefore any evidence that an advanced transductive inference for computers exists is very important for understanding the nature of human reason.

This book was almost finished when I realized that it would not be easy for a reader to discern the general philosophy (which is nontrivial) that ties together many technical details (some of which are very sophisticated). Therefore, I decided to stop working on this book for a while and to write a short, simplified version that would contain neither proofs nor unnecessary technical details, but would contain informal reasoning and comments. In 1995, I published that book.¹

Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*, Springer-Verlag, 1995, 189 pages.

In contrast to the short one, the present book contains the proofs of all the main assertions. Nevertheless, it is not merely a collection of proofs of the statements described in *The Nature of Statistical Learning Theory*. More details of the theory made it possible to display more deeply both the details of the philosophy and the details of the new methods.

The three years between completion of the short book and this one were very fruitful in developing **SRM** methodology. During this time, new methods of function estimation in multidimensional spaces based on the **SVM** techniques were developed. These methods go beyond learning theory. They can be considered as a general approach to function representation in high-dimensional spaces that in many cases can overcome the "curse of dimensionality." The details of these methods are described in the book.

As with the short one, this book is devoted to an approach that in many respects differs from classical statistical approaches. One can consider it as an attempt to create a new paradigm in statistics that depends less on heuristics and instead is connected to the inductive theory of inference.

It is my hope that the book displays how deeply learning theory is connected to both induction theory and the fundamentals of statistics and how these connections give rise to effective practical methods of inference.



THEORY OF LEARNING AND GENERALIZATION

Part I analyses factors responsible for generalization and shows how to control these factors in order to generalize well.

TWO APPROACHES TO THE LEARNING PROBLEM

In this chapter we consider two approaches to the learning problem—the problem of choosing the desired dependence on the basis of empirical data.

The first approach is based on the idea that the quality of the chosen function can be evaluated by a risk functional. In this case the choice of the approximating function from a given set of functions is a problem of minimizing the risk functional on the basis of empirical data. This problem is rather general. It embeds many problems of statistics. In this book we consider three of them: pattern recognition, regression estimation, and density estimation.

The second approach to the learning problem is based on estimating dcsircd stochastic dcpndcncics (dcnsitics, conditional densities, conditional probabilities). It requires solution of integral equations (determining these dependencies) in situations where some elements of the equations are known only approximately. Using estimated stochastic dependence, the pattern recognition and regression estimation problems can be solved as well. However, the function obtained by solution of the integral equations provides much more details than is required for these problems. The price we pay for these details is the necessity to solve ill-posed problems.

1.1 GENERAL MODEL OF LEARNING FROM EXAMPLES

Consider the following model of searching for functional dependency, which we call the *model of learning from examples*.

The model contains three elements (Fig 1.1):

1. The generator of the data (examples), G.

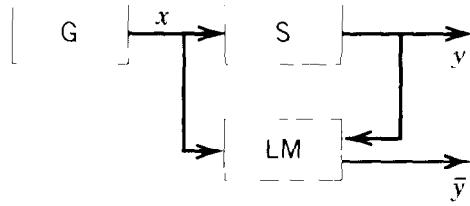


FIGURE 1.1. A model of learning from examples. During the learning process, the learning machine observes the pairs (x, y) (the training set). After training, the machine must on any given x return a value \bar{y} . The goal is to return a value y which is close to the supervisor's response y .

2. The target operator (sometimes called *supervisor's operator* or, for simplicity, *supervisor*), S .
3. The learning machine. LM .

The generator G is a source of situations that determines the environment in which the supervisor and the learning machine act. In this book, we consider the simplest environment: G generates the vectors $x \in X$ *independently and identically distributed* (i.i.d.) according to some unknown (but fixed) probability distribution function $F(x)$.

These vectors are inputs to the target operator (supervisor): the target operator returns the output values y . The target operator, which transforms the vectors x into values y , is unknown, but we know that it exists and does not change.

The learning machine observes I' pairs

$$(x_1, y_1), \dots, (x_{I'}, y_{I'})$$

(the *training set*) which contain input vectors x and the supervisor's response y . During this period, the learning machine constructs some operator which will be used for prediction of the supervisor's answer y_i on any specific vector x_i generated by the generator G . The goal of the learning machine is to construct an appropriate approximation.

To be a mathematical statement, this general scheme of learning from examples needs some clarification. First of all, we have to describe what kind of operators are used by the supervisor. In this book, we suppose that the supervisor returns the output y on the vector x according to a conditional distribution function $F(y|x)$ (this includes the case when the supervisor uses some function $y = f(x)$).

Thus, the learning machine observes the training set, which is drawn randomly and independently according to a joint distribution function $F(x,y) = F(x)F(y|x)$. (Recall that we do not know this function but we do know that it exists.) Using this training set, the learning machine constructs an approximation to the unknown operator.

To construct an approximation, the learning machine chooses one of the

two goals to pursue:

- To *imitate* the supervisor's operator: Try to construct an operator which provides for a given generator G , the best prediction to the supervisor's outputs.
- To *identify* the supervisor's operator: Try to construct an operator which is close to the supervisor's operator.

There exists an essential difference in these two goals. In the first case, the goal is to achieve the best results in prediction of the supervisor's outputs for the environment given by the generator G . In the second case, to get good results in prediction is not enough; it is required to construct an operator which is close to the supervisor's one in a given metric. These two goals of the learning machine imply two different approaches to the learning problem.

In this book we consider both approaches. We show that the problem of imitation of the target operator is easier to solve. For this problem, a nonasymptotic theory will be developed. The problem of identification is more difficult. It refers to the so-called ill-posed problems. For these problems, only an asymptotic theory can be developed. Nevertheless, we show that the solutions for both problems are based on the same general principles.

Before proceeding with the formal discussion of the learning problem, we have to make a remark. We have to explain what it means "to construct an operator" during the learning process. From a formal point of view, this means that the learning machine can implement some fixed set of functions given by the construction of the machine. During the learning process, it chooses from this set an appropriate function. The rule for choosing the function is one of the most important subjects of the theory and it will be discussed in this book. But the general assertion is:

The learning process is a process of choosing an appropriate function from a given set of functions.

We start our discussion of the learning problem with the problem of imitation. It is based on the general statistical problem of minimizing the risk functional on the basis of empirical data. In the next section we consider a statement of this problem, and then in the following sections we demonstrate that different learning problems are particular cases of this general one.

1.2 THE PROBLEM OF MINIMIZING THE RISK FUNCTIONAL FROM EMPIRICAL DATA

Each time the problem of selecting a function with desired quality arises, the same model may be considered: Among the totality of possible functions, one

looks for the one that satisfies the given quality criterion in the best possible manner.

Formally this means that on the subset Z of the vector space H^n , a set of admissible functions $\{g(z)\}$, $z \in Z$, is given, and a functional

$$R = R(g(z)) \quad (1.1)$$

is defined which is the criterion of quality of the chosen function. It is then required to find the function $g^*(z)$ from the set $\{g(z)\}$ which minimizes the functional (1.1). (We shall assume that the minimum of the functional corresponds to the best quality and that the minimum of (1.1) exists in $\{g(z)\}$.) In the case when the set of functions $\{g(z)\}$ and the functional $R(g(z))$ are explicitly given, the search for the function $g^*(z)$ which minimizes $R(g(z))$ is the subject of the calculus of variations.

In this book, another case is considered, when a probability distribution function $F(z)$ is defined on Z and the functional is defined as the mathematical expectation

$$R(g(z)) = \int L(z, g(z)) dF(z), \quad (1.2)$$

where function $L(z, g(z))$ is integrable for any $g(z) \in \{g(z)\}$. The problem is to minimize the functional (1.2) in the case when the probability distribution $F(z)$ is unknown but the sample

$$z_1, \dots, z_s \quad (1.3)$$

of observations drawn randomly and independently according to $F(z)$ is available.

Sections 1.3, 1.4, and 1.5 shall verify that the basic statistical problems related to function estimation problem can be reduced to the minimization of (1.2) based on empirical data (1.3). Meanwhile, we shall note that there is a substantial difference between problems arising when the functional (1.1) is minimized directly and those encountered when the functional (1.2) is minimized on the basis of empirical data (1.3).

In the case of minimizing (1.1), the problem is to organize the search for a function $g^*(z)$ from the set $\{g(z)\}$ which minimizes (1.1). When (1.2) is to be minimized on the basis of empirical data (1.3), the basic problem is to formulate a constructive criterion for choosing the function rather than organizing the search of the functions in $\{g(z)\}$. (The functional (1.2) by itself cannot serve as a selection criterion, since the measure $F(z)$ involved in it is unknown.) Thus, in the first case, the question is:

How can we obtain the minimum of the functional in the given set of functions?

While in the second case the question is:

What should be minimized in order to select from the set $\{g(z)\}$ a function which will guarantee that the functional (1.2) is small?

Strictly speaking, one cannot minimize (1.2) based on (1.3) using methods developed in optimization theory. The minimization of the functional (1.2) on the basis of empirical data (1.3) is one of the main problems of mathematical statistics.

When formulating the minimization problem for functional (1.2), the set of functions $g(z)$ will be given in a parametric form $\{g(z, \alpha), \alpha \in \Lambda\}$.[†] Here α is a parameter from the set A such that the value $\alpha = \alpha^*$ defines the specific function $g(z, \alpha^*)$ in the set $g(z, \alpha)$. Finding the required function means determining the corresponding value of the parameter $\alpha \in A$.

The study of only parametric sets of functions is not a restriction on the problem, since the set A , to which the parameter α belongs, is arbitrary: It can be a set of scalar quantities, a set of vectors, or a set of abstract elements.

In the new notation the functional (1.2) can be rewritten as

$$R(\alpha) = \int Q(z, \alpha) dF(z), \quad \alpha \in \Lambda, \quad (1.4)$$

where

$$Q(z, \alpha) = L(z, g(z, \alpha)).$$

The function $Q(z, \alpha)$, which depends on two variables z and α , is called the *loss function*.

The problem of minimizing functional (1.4) admits a simple interpretation: It is assumed that each function $Q(z, \alpha^*)$, $\alpha^* \in A$ (i.e., each function of z for a fixed $\alpha = \alpha^*$), determines the amount of the *loss* resulting from the realization of the vector z . The *expected loss* (with respect to z) for the function $Q(z, \alpha^*)$ is determined by the integral

$$R(\alpha^*) = \int Q(z, \alpha^*) dF(z)$$

This functional is called the *risk functional* or the *risk*. The problem is to choose in the set $Q(z, \alpha)$, $\alpha \in A$, a function $Q(z, \alpha_0)$ which minimizes the risk when the probability distribution function is unknown but random independent observations z_1, \dots, z_ℓ are given.

Remark. Let us clarify the phrase "probability distribution function is unknown." Denote by \mathcal{P}_0 the set of all possible probability distribution functions on Z and by \mathcal{P} some subset of probability distribution functions from \mathcal{P}_0 .

[†]We shall always omit the braces when writing a set of functions. A single function is distinguished from a set of functions by indicating whether the parameter α is fixed or not.

We will distinguish between two cases:

1. Case where we have no information about the unknown distribution function. (We have only the trivial information that $F(z) \in \mathcal{P}_0$.)
2. Case where we have nontrivial information about the unknown distribution function. We know that $F(z)$ belongs to the subset \mathcal{P} which does not coincide with \mathcal{P}_0 .

In this book, we consider mostly the first case, where we have no a priori information about the unknown distribution function. However, we will consider the general method for constructing a theory which is valid for any given set of probability measures.

The problem of minimizing the risk functional (1.4) on the basis of empirical data (1.3) is rather general. It includes in particular three basic statistical problems:

1. The problem of pattern recognition
2. The problem of regression estimation
3. The problem of density estimation

In the next sections we shall verify that all these problems can be reduced to the minimization of the risk functional (1.4) on the basis of the empirical data (1.3).

1.3 THE PROBLEM OF PATTERN RECOGNITION

The *problem of pattern recognition* was formulated in the late 1950s. In essence it can be stated as follows: A supervisor observes occurring situations and determines to which of k classes each one of them belongs. It is required to construct a machine which, after observing the supervisor's classification, carries out the classification approximately in the same manner as the supervisor.

Using formal language, this statement can be expressed as follows: In a certain environment characterized by a probability distribution function $F(x)$, situation x appears randomly and independently. The supervisor classifies each situations into one of k classes. We assume that the supervisor carries out this classification using the conditional probability distribution function $F(\omega|x)$, where $\omega \in \{0, 1, \dots, k - 1\}$ ($\omega = p$ indicates that the supervisor assigns situation x to the class number p).[†]

[†]This is the most general case which includes a case when a supervisor classifies situations x using a function $\omega = f(x)$.

Neither the properties of the environment $F(x)$ nor the decision rule of the supervisor $F(\omega|x)$ are known. However, we do know that both functions exist. Thus, a joint distribution $F(\omega, x) = F(\omega|x)F(x)$ exists.

Now, let a set of functions $\phi(x, a)$, $a \in A$, which take only k values $\{0, 1, \dots, k - 1\}$ (a set of decision rules), be given. We shall consider the simplest loss function

$$L(\omega, \phi) = \begin{cases} 0 & \text{if } \omega = \phi \\ 1 & \text{if } \omega \neq \phi \end{cases} \quad (1.5)$$

The problem of pattern recognition is to minimize the functional

$$R(\alpha) = \int L(\omega, \phi(x, \alpha)) dF(\omega, x) \quad (1.6)$$

on the set of functions $\phi(x, a)$, $a \in A$, where the distribution function $F(\omega, x)$ is unknown but a random independent sample of pairs

$$(\omega_1, x_1), \dots, (\omega_\ell, x_\ell) \quad (1.7)$$

is given. For the loss function (1.5), the functional (1.6) determines the probability of a classification error for any given decision rule $\phi(x, a)$.

The problem, therefore, is to minimize the probability of a classification error when the probability distribution function $F(\omega, x)$ is unknown but the data (1.7) are given.

For simplicity consider the two-class classification problem (i.e., $\omega \in \{0, 1\}$) where we use the simplest loss function (1.5).

Thus, the problem of pattern recognition has been reduced to the problem of minimizing the risk on the basis of empirical data. The special feature of this problem is that the set of loss functions $Q(z, a)$, $a \in A$, is not arbitrary as in the general case described in Section 1.2. The following restrictions are imposed:

- The vector z consists of $n+1$ coordinates: coordinate ω , which takes on only a finite number of values (two values for a two classes problem), and n coordinates x^1, \dots, x^n which form the vector x .
- The set of functions $Q(z, a)$, $a \in A$, is given by

$$Q(z, \alpha) = L(\omega, \phi(x, \alpha)), \quad \alpha \in \Lambda$$

and also takes on only a finite number of values (zero and one for the simplest loss function).

This specific feature of the risk minimization problem characterizes the pattern recognition problem. The problem of pattern recognition forms the simplest learning problem because it deals with the simplest loss function. The loss function in the pattern recognition problem describes a set of indicator functions—that is, functions that take only two values, zero and one.

1.4 THE PROBLEM OF REGRESSION ESTIMATION

Two sets of elements X and Y are connected by a functional dependence if to each element $x \in X$ there corresponds a unique element $y \in Y$. This relationship is called a function if X is a set of vectors and Y is a set of scalars.

However, there exist relationships (stochastic dependencies) where to each vector x there corresponds a number y which we obtain as a result of random trials. For each x , let a distribution $F(y|x)$ be defined on Y according to which the selection of the value of y is implemented. The function of the conditional probability expresses the stochastic relationship between y and x .

Now, let the vectors x appear randomly and independently in accordance with a distribution $F(x)$. Then, in accordance with $F(y|x)$, the values of y are realized in random trials. In this case, there exists a joint distribution function $F(x, y)$. In accordance with this measure the observed pairs

$$(y_1, x_1), \dots, (y_t, x_t)$$

are formed randomly and independently. Estimating the stochastic dependence based on this empirical data means estimating the conditional distribution function $F(y|x)$, and this is indeed quite a difficult problem. As we show, it leads to the need to solve so-called ill-posed problems.

However, the knowledge of the function $F(y|x)$ is often not required; it is sufficient to determine one of its characteristics, for example the function of conditional mathematical expectation:

$$r(x) = \int y dF(y|x). \quad (1.8)$$

This function is called the *regression*, and the problem of its estimation in the set of functions $f(x, \alpha)$, $\alpha \in \Lambda$, is referred to as the problem of regression estimation. We now show that under conditions

$$\int y^2 dF(y|x) < \infty, \quad \int r^2(x) dF(y|x) < \infty$$

the problem of regression estimation is reduced to the model of minimizing risk based on empirical data.

Indeed, on the set $f(x, \alpha)$, $\alpha \in \Lambda$ ($f(x, \alpha) \in L_2(P)$), the minimum of the functional

$$R(\alpha) = \int (y - f(x, \alpha))^2 dF(y|x) \quad (1.9)$$

(provided the minimum exists) is attained at the regression function if the regression $r(x)$ belongs to $f(x, n)$, $\alpha \in \Lambda$. The minimum of this functional is attained at the function $f(x, \alpha^*)$, which is the closest to regression $r(x)$ in the

metric $L_2(P)$

$$\rho(f_1, f_2) = \sqrt{\int (f_1(x) - f_2(x))^2 dF(x)}$$

if the regression $r(x)$ does not belong to the set $f(x, \alpha), \alpha \in A$.

To show this, denote

$$\Delta f(x, \alpha) = f(x, \alpha) - r(x).$$

Then functional (1.9) can be written in the form

$$\begin{aligned} R(\alpha) &= \int (y - r(x))^2 dF(y, x) + \int (\Delta f(x, \alpha))^2 dF(y, x) \\ &\quad - 2 \int \Delta f(x, \alpha)(y - r(x)) dF(y, x). \end{aligned}$$

In this expression, the third summand is zero, since according to (1.8)

$$\begin{aligned} &\int \Delta f(x, \alpha)(y - r(x)) dF(y, x) \\ &= \int \Delta f(x, \alpha) \left[\int (y - r(x)) dF(y|x) \right] dF(x) = 0. \end{aligned}$$

Thus we have verified that

$$R(\alpha) = \int (y - r(x))^2 dF(y, x) + \int (f(x, \alpha) - r(x))^2 dF(x).$$

Since the first summand does not depend on α , the function $f(x, \alpha_0)$, which minimizes the risk functional $R(\alpha)$, is the regression if $r(x) \in f(x, \alpha)$, or the function $f(x, \alpha_0)$ which minimizes the risk functional $R(\alpha)$ is the closest function to the regression (in the metric $L_2(P)$), if $r(x)$ does not belong to $f(x, \alpha)$.

This equation also implies that if the regression function $r(x) = f(x, \alpha_0)$ belongs to the given set of functions $f(x, \alpha), \alpha \in A$, and if for some function $f(x, \alpha^*)$ the risk functional $R(\alpha^*)$ is ε -close to the minimal one

$$R(\alpha^*) - \inf_{\alpha \in A} R(\alpha) < \varepsilon,$$

then the function $f(x, \alpha^*)$ is $\sqrt{\varepsilon}$ -close to the regression in the metric $L_2(P)$:

$$\rho(f(x, \alpha^*), r(x)) = \sqrt{\int (f(x, \alpha^*) - r(x))^2 dF(x)} < \sqrt{\varepsilon}.$$

Thus, the problem of estimating the regression may be also reduced to the scheme of minimizing expected risk. The specific feature of this problem is that the set of functions $Q(z, \alpha)$, $\alpha \in \Lambda$, is subject to the following restrictions:

- The vector z consists of $n+1$ coordinates: the coordinate y and n coordinates x^1, \dots, x^n forming the vector x . However, in contrast to the pattern recognition problem, the coordinate y as well as the function $f(x, \alpha)$ may take any value in the interval $(-\infty, \infty)$
- The set of loss functions $Q(z, \alpha)$, $\alpha \in \Lambda$, is of the form

$$Q(z, \alpha) = (y - f(x, \alpha))^2.$$

The important feature of the regression estimation problem is that the loss-function $Q(z, \alpha)$ can take on arbitrary non-negative values whereas in pattern recognition problem it can take only two values.

1.5 PROBLEM OF INTERPRETING RESULTS OF INDIRECT MEASURING

Along with the problem of regression estimation we consider the problem of estimating functional dependencies from indirect measuring.

Suppose one would like to estimate a function $f(t)$ that can be measured at no point of t . At the same time, another function $F(x)$ which is connected with $f(t)$ by operator

$$Af(t) = F(x)$$

may admit measurements. It is then required on the basis of measurements (with errors ξ)

$$y_1, \dots, y_t, \quad y_i = F(x_i) + \xi_i$$

of function $F(x)$ at points x_1, \dots, x_t to obtain in a set $f(t, \alpha)$ the solution of the equation. This problem is called the problem of *interpreting results of indirect measurements*.

The formation of the problem is as follows: Given a continuous operator A which maps in one-to-one manner the elements $f(t, \alpha)$ of a metric space E_1 into the elements $F(x, \alpha)$ of a metric space E_2 , it is required to obtain a solution of the operator equation in a set of functions $f(t, \alpha)$, $\alpha \in \Lambda$, provided that the function $F(x)$ is unknown, but measurements y_1, \dots, y_t are given.

We assume that the measuring $F(x)$ does not involve systematic error. that is,

$$Ey_i = F(x_i)$$

and the random variables y_{x_i} and y_{x_j} ($i \neq j$) are independent. We also assume that function is defined on the interval $[a,b]$. The points x at which measurements of the function $F(x)$ are carried out are randomly and independently distributed on $[a,b]$ according to uniform distribution.[†]

The problem of interpreting results of indirect experiments also can be reduced to the problem of minimizing the expected risk based on empirical data. Indeed, consider the functional

$$R(\alpha) = \int (y - Af(t, \alpha))^2 p(y|x) dy dx$$

Using the same decomposition technique as in the previous section we obtain

$$\begin{aligned} R(\alpha) &= \int (y - F(x, \alpha))^2 p(y|x) dy dx \\ &= \int (y - Af(t))^2 p(y|x) dy dx + \int (F(x, \alpha) - F(x))^2 dx \end{aligned}$$

where $f(t)$ and $F(x)$ are the solution of integral equation and its image in E_2 space.

We have thus again arrived at setup for minimizing expected risk on the basis of empirical data. To solve this problem, we have to find function $f(t, \alpha_0)$, the image of which is the regression function in E_2 space.

- The vector z consists of $n+1$ coordinates: the coordinate y and n coordinates x', \dots, x'' forming the vector x .
- The set of loss-functions $Q(z, \alpha)$, $\alpha \in A$, is of the form

$$Q(z, \alpha) = (y - Af(t, \alpha))^2.$$

The specific feature of interpreting results of indirect experiments that the problem of solving operator equation

$$Af(t) = F(x), \quad f(t) \in f(t, \alpha)$$

may be ill-posed (we will discuss this problem below). In this case not all good approximations to the regression $F(x)$ imply good approximations to the desired solution $f(t)$. In order to approximate the solution of the operator equation well, one has to choose the function that not only provides a small value to the risk functional, but also satisfies some additional constraints that we will discuss later.

[†] The points x can be defined by any nonvanishing density on $[a,b]$.

1.6 THE PROBLEM OF DENSITY ESTIMATION (THE FISHER-WALD SETTING)

Let $p(x, \alpha), \alpha \in \Lambda$, be a set of probability densities containing the required density

$$p(x, \alpha_0) = \frac{dF(x)}{dx}.$$

Consider the functional

$$R(\alpha) = - \int \ln p(x, \alpha) dF(x). \quad (1.10)$$

Below we show that:

1. The minimum of the functional (1.10) (if it exists) is attained at the functions $p(x, \alpha^*)$ which may differ from $p(x, \alpha_0)$ only on a set of zero measure.
2. 'The Bretagnolle-Huber inequality'

$$\int |p(x, \alpha) - p(x, \alpha_0)| dx \leq 2\sqrt{1 - \exp\{R(\alpha_0) - R(\alpha)\}} \quad (1.11)$$

is valid.

Therefore, the functions $p(x, \alpha^*)$ which are ε -close to the minimum

$$R(\alpha^*) - \inf_{\alpha \in \Lambda} R(\alpha) < \varepsilon$$

will be $2\sqrt{1 - \exp\{-\varepsilon\}}$ -close to the required density in the metric L_1 .

The proof of the first assertion is based on the *Jensen inequality*, which states that for a concave function ψ the inequality

$$\int \psi(\Phi(x)) dF(x) \leq \psi \left(\int \Phi(x) dF(x) \right) \quad (1.12)$$

is valid.

Consider the functions

$$\psi(u) = \ln u, \quad \Phi(x) = \frac{p(x, \alpha)}{p(x, \alpha_0)}.$$

Jensen's inequality implies

$$\int \ln \frac{p(x, \alpha)}{p(x, \alpha_0)} dF(x) \leq \ln \int \frac{p(x, \alpha)}{p(x, \alpha_0)} p(x, \alpha_0) dx = \ln 1 = 0.$$

So, the inequality

$$\int \ln \frac{p(x, \alpha)}{p(x, \alpha_0)} dF(x) = \int \ln p(x, \alpha) dF(x) - \int \ln p(x, \alpha_0) dF(x) \leq 0$$

is valid. Taking into account the sign in front of the integral (1.10), this inequality proves our first assertion.

To prove the Bretagnolle–Huber inequality, use the following identity:

$$\begin{aligned} \int p(x, \alpha_0) \ln \frac{p(x, \alpha)}{p(x, \alpha_0)} dx &= \int p(x, \alpha_0) \ln \left[\min \left(\frac{p(x, \alpha)}{p(x, \alpha_0)}, 1 \right) \right] dx \\ &\quad + \int p(x, \alpha_0) \ln \left[\max \left(\frac{p(x, \alpha)}{p(x, \alpha_0)}, 1 \right) \right] dx. \end{aligned}$$

We apply Jensen's inequality to both terms on the right-hand side of this equality

$$\begin{aligned} \int p(x, \alpha_0) \ln \frac{p(x, \alpha)}{p(x, \alpha_0)} dx &\leq \ln \int \min(p(x, \alpha), p(x, \alpha_0)) dx \\ &\quad + \ln \int \max(p(x, \alpha), p(x, \alpha_0)) dx. \end{aligned} \quad (1.13)$$

Note that the following identities are true:

$$\begin{aligned} \min(a, b) &= \frac{a + b - |a - b|}{2}, \\ \max(a, b) &= \frac{a + b + |a - b|}{2}. \end{aligned} \quad (1.14)$$

Substituting (1.14) into (1.13), we obtain

$$\begin{aligned} &\int p(x, \alpha_0) \ln \frac{p(x, \alpha)}{p(x, \alpha_0)} dx \\ &\leq \ln \left\{ \left(1 - \frac{1}{2} \int |p(x, \alpha) - p(x, \alpha_0)| dx \right) \left(1 + \frac{1}{2} \int |p(x, \alpha) - p(x, \alpha_0)| dx \right) \right\} \\ &= \ln \left(1 - \left(\frac{1}{2} \int |p(x, \alpha) - p(x, \alpha_0)| dx \right)^2 \right). \end{aligned} \quad (1.15)$$

This inequality implies Bretagnolle–Huber inequality.

Thus, the problem of estimating the density in L_1 is reduced to the minimization of the functional (1.10) on the basis of empirical data. We call this setting of the density estimation problem the Fisher–Wald's setting. (In Section 1.8 we consider another setting of this problem.)

The special feature of the density estimation problem in the Fisher–Wald setting is that the set of functions $\mathcal{Q}(z, a)$ is subject to the following restrictions:

- The vector z coincides with the vector x .
- The set of functions $Q(z, \alpha), \alpha \in \Lambda$, is of the form

$$Q(z, \alpha) = -\log p(x, \alpha),$$

where $p(x, \alpha)$ is a set of density functions. The loss function $Q(z, \alpha)$ takes on arbitrary values on the interval $(-\infty, \infty)$, whereas in the regression estimation problem it takes on only nonnegative values.

We will restrict our analysis to these three problems. However, many other problems of estimating empirical dependencies can be reduced to the model of risk minimization based on empirical data.

1.7 INDUCTION PRINCIPLES FOR MINIMIZING THE RISK FUNCTIONAL ON THE BASIS OF EMPIRICAL DATA

In the previous sections, we considered the problem of minimizing the risk functional on the basis of empirical data. It was shown that different problems such as pattern recognition, regression estimation, and density estimation can be reduced to this scheme by specifying a loss function in the risk functional.

Now a main question arises:

How can we minimize the risk functional?

We cannot minimize the functional directly since the probability distribution function $F(x)$ that defines the risk is unknown. What shall we do instead? The answer to this question determines an *induction principle* for solving learning problems.

In this book, two induction principles will be considered: (1) the classical one which we introduce in this section and (2) a new one which we consider in Chapter 6.

Principle of Empirical Risk Minimization. Let us, instead of minimizing the risk functional

$$R(\alpha) = \int Q(z, \alpha) dF(z), \quad \alpha \in \Lambda,$$

minimize the functional

$$R_{\text{emp}}(\alpha) = \frac{1}{t} \sum_{i=1}^t Q(z_i, \alpha), \quad \alpha \in \Lambda, \quad (1.16)$$

which we call the empirical risk functional. The empirical risk functional is constructed on the basis of data

$$z_1, \dots, z_\ell$$

obtained according to distribution function $F(z)$. This functional is defined in explicit form, and it is subject to minimization.

Let the minimum of the risk functional be attained at $Q(z, \alpha_0)$ and let the minimum of the empirical risk functional be attained at $Q(z, \alpha_\ell)$. We shall consider the function $Q(z, \alpha_\ell)$ as an approximation to the function $Q(z, \alpha_0)$. This principle of solving the risk minimization problem is called the *empirical* risk minimization (induction) principle.

The study of this principle is one of the main subjects of this book. The problem is to establish conditions under which the obtained function $Q(z, \alpha_\ell)$ is close to the desired one, $Q(z, \alpha_0)$.

1.8 CLASSICAL METHODS FOR SOLVING FUNCTION ESTIMATION PROBLEMS

Below we show that classical methods for solving our three statistical problems (pattern recognition, regression estimation, and density estimation) are implementations of the principle of empirical risk minimization.

Method of Minimizing Number of Training Error. In Section 1.3 we showed that the minimization using empirical data (training data)

$$(\omega_1, x_1), \dots, (\omega_\ell, x_\ell)$$

of the risk functional

$$R(\alpha) = \int L(\omega, \phi(x, \alpha)) dF(\omega, x), \quad \alpha \in \Lambda$$

on a set of functions $\phi(x, a), a \in A$, that take on only a finite number of values renders the pattern recognition problem.

Consider the empirical risk functional

$$R_{\text{emp}}(\alpha) = \frac{1}{\ell} \sum_{i=1}^{\ell} L(\omega_i, \phi(x_i, \alpha)), \quad \alpha \in \Lambda.$$

In the case when $L(\omega_i, \phi) \in \{0, 1\}$ (0 if $\omega = \phi$ and 1 if $\omega \neq \phi$), minimization of the empirical risk functional produced a function which has the smallest number of errors on the training data.

Least Squares Method for the Regression Estimation Problem. In Section 1.4. we considered the problem of regression estimation as the problem of minimization of the functional

$$R(\alpha) = \int (y - f(x, \alpha))^2 dF(y, x), \quad \alpha \in \Lambda$$

on the set of functions $f(x, \alpha)$, $\alpha \in \Lambda$, on the basis of empirical data

$$(y_1, x_1), \dots, (y_t, x_t)$$

For this functional, the empirical risk functional is

$$R_{\text{emp}}(\alpha) = \frac{1}{t} \sum_{i=1}^t (y_i - f(x_i, \alpha))^2, \quad \alpha \in \Lambda.$$

According to the empirical risk minimization principle. to estimate the regression function we have to minimize this functional. In statistics, the method of minimizing this functional is known as the "least-squares method."

Maximum Likelihood Method for Density Estimation. In Section 1.5, we considered the problem of density estimation as the problem of minimization of the functional

$$R(\alpha) = - \int \ln p(x, \alpha) dF(x), \quad \alpha \in \Lambda$$

on the set of densities $p(x, \alpha)$, $\alpha \in \Lambda$, using independent identically distributed data

$$x_1, \dots, x_t.$$

For this functional, the empirical risk functional is

$$R_{\text{emp}}(\alpha) = - \sum_{i=1}^t \ln p(x_i, \alpha).$$

According to the principle of empirical risk minimization. the minimum of this functional provides an approximation of the density. It is the same solution which comes from the maximum likelihood method. (In the maximum likelihood method. a plus sign is used in front of the sum instead of a minus.)

Thus, we find that the classical methods of solving our statistical problems are realizations of the general induction principle of minimizing empirical risk. In subsequent chapters. we will study the general methods of minimizing the risk functionals and then apply them to our specific problems. But before that. we will consider a second approach to the learning problems, which is not based on the scheme of minimizing the risk functional from empirical data.

1.9 IDENTIFICATION OF STOCHASTIC OBJECTS: ESTIMATION OF THE DENSITIES AND CONDITIONAL DENSITIES

1.9.1 Problem of Density Estimation. Direct Setting

Consider methods for identifying stochastic objects. We start with the problem of density estimation. Let ξ be a random variable. The probability of random event

$$F(x) = P\{\xi < x\}$$

is called a *probability distribution function* of the random variable ξ . A random vector $\bar{\xi}$ is a generalization of the notion of a random variable. The function

$$F(x) = P\{\bar{\xi} < \bar{x}\},$$

where the inequality is interpreted coordinatewise, is called a *probability distribution function of the random vector $\bar{\xi}$* .

We say that the random variable ξ (random vector $\bar{\xi}$) has a density if there exists a nonnegative function $p(u)$ such that for all x the equality

$$F(x) = \int_{-\infty}^x p(u) du$$

is valid.

The function $p(x)$ is called a *probability density* of the random variable (random vector). So, by definition, to estimate a probability density from the data we need to obtain a solution of the integral equation

$$\int_{-\infty}^x p(u, a) du = F(x) \quad (1.17)$$

on a given set of densities $p(x, a)$, $a \in A$, under conditions that the distribution function $F(x)$ is unknown and a random independent sample

$$x_1, \dots, x_\ell, \quad (1.18)$$

obtained in accordance with $F(x)$, is given.

One can construct approximations to the distribution function $F(x)$ using the data (1.18)—for example, the so-called *empirical distribution function* (1.18) (see Fig. 1.2):

$$F_\ell(x) = \frac{1}{\ell} \sum_{i=1}^{\ell} \theta(x - x_i), \quad (1.19)$$

where we define for vector[†] u the step function

$$\theta(u) = \begin{cases} 1 & \text{all coordinates of the vector } u \text{ are positive.} \\ 0 & \text{otherwise.} \end{cases}$$

[†] Including scalars as one-dimensional vectors.

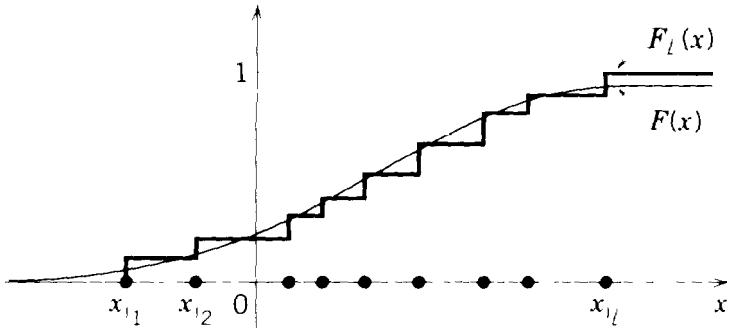


FIGURE 1.2. The empirical distribution function $F_l(x)$, constructed from the data x_1, \dots, x_l , approximates the probability distribution function $F(x)$.

In the next section, we will show that empirical distribution function $F_l(x)$ is a *good* approximation to the actual distribution function $F(x)$.

Thus, the problem of density estimation is to find an approximation to the solution of the integral equation (1.17) if the probability distribution function is unknown; however, an approximation to this function can be defined.

We call this setting of the density estimation problem *direct setting* because it based on the definition of density. In the following sections we shall discuss the problem of solving integral equations with an approximate right-hand side, but now we turn to a direct setting of the problem of estimating the conditional probability. Using the conditional probability, one can easily solve the pattern recognition problem.

1.9.2 Problem of Conditional Probability Estimation

Consider pairs (ω, x) , where x is a vector and ω is a scalar which takes on only k values $\{0, 1, \dots, k - 1\}$. According to the definition, the conditional probability $P(\omega|x)$ is a solution of the integral equation

$$\int_{-\infty}^x P(\omega|t) dF(t) = F(\omega, x), \quad (1.20)$$

where $F(x)$ is the distribution function of random vectors x , and $F(\omega, x)$ is the joint distribution function of pairs (ω, x) .

The problem of estimating conditional probability in the set of functions $P_\alpha(\omega|x)$, $\alpha \in \Lambda$, is to obtain an approximation to the solution of the integral equation (1.20) when both distribution functions $F(x)$ and $F(\omega, x)$ are unknown but the data

$$(\omega_1, x_1), \dots, (\omega_\ell, x_\ell)$$

is given. As in the case of density estimation, we can approximate the unknown distribution functions $F(x)$ and $F(\omega, x)$ by the empirical distribution

functions (1.19) and function

$$F_\ell(\omega, x) = \frac{1}{\ell} \sum_{i=1}^{\ell} \theta(x - x_i) \delta(\omega, x_i),$$

where

$$\delta(\omega, x) = \begin{cases} 1 & \text{if the vector } x \text{ belongs to the class } \omega, \\ 0 & \text{otherwise.} \end{cases}$$

Thus, the problem is to obtain an approximation to the solution of integral equation (1.20) in the set of functions $P_a(\omega|x)$, $a \in A$, when probability distribution functions $F(x)$ and $F(\omega, x)$ are unknown, but approximations $F_\ell(x)$ and $F_\ell(\omega, x)$ are given.

Note that estimation of the conditional probability function $F(\omega|x)$ is a stronger solution to the pattern recognition problem than the one considered in Section 1.3. In Section 1.3, the goal was to find the best decision rule from the *given set of decision rules*; it did not matter whether this set did or did not contain a good approximation to the supervisor's decision rule. In this statement of the identification problem, the goal is to find the best approximation to the supervisor's decision rule (which is the conditional probability function according to the statement of the problem). Of course, if the supervisor's operator $F(\omega|x)$ is known, then one can easily construct the optimal decision rule. For the case where $\omega \in \{0, 1\}$ and a priori probability of classes are equal, it has the form

$$f(x) = \theta(P(\omega = 1|x) - \frac{1}{2}).$$

This is the so-called Bayes rule; it assigns vector x to class 1 if the probability that this vector belongs to the first class is larger than 1/2 and assigns 0 otherwise. However, the knowledge of the conditional probability not only gives the best solution to the pattern recognition problem, but also provides an estimate of the error probability for any specific vector x .

1.9.3 Problem of Conditional Density Estimation

Finally, consider the problem of conditional density estimation. In the pairs (y, x) , let the variables y be scalars and let x be vectors. Consider the equality

$$\int_{-\infty}^y \int_{-\infty}^x p(t|u) dF(u) dt = F(y, x), \quad (1.21)$$

where $F(x)$ is a probability distribution function which has a density, $p(y|x)$

is the conditional density of y given x , and $F(y, x)$ is the joint probability distribution function[†] defined on the pairs (y, x) .

As before, we are looking for an approximation to the conditional density $p(y|x)$ by solving the integral equation (1.21) on the given set of functions when both distribution functions $F(x)$ and $F(y, x)$ are unknown: and the random, i.i.d. pairs

$$(y_1, x_1), \dots, (y_\ell, x_\ell) \quad (1.22)$$

are given. As before, we can approximate the empirical distribution function $F_t(x)$ and empirical distribution function

$$F_t(y, x) = \frac{1}{\ell} \sum_{i=1}^{\ell} \theta(y - y_i) \theta(x - x_i).$$

Thus, our problem is to get an approximation to the solution of the integral equation (1.21) in the set of functions $p_a(y|x)$, $a \in \Lambda$, when the probability distribution functions are unknown but we can construct the approximations $F_t(x)$ and $F_t(y, x)$ using data (1.22).

Note that the conditional density $p(y|x)$ contains much more information about the behavior of the random value y for fixed x than the regression function. The regression function can be easily obtained from conditional density (see the definition of the regression function (1.8)).

1.10 THE PROBLEM OF SOLVING AN APPROXIMATELY DETERMINED INTEGRAL EQUATION

All three problems of stochastic dependencies estimation can be described in the following general way. It is necessary to solve a linear continuous operator equation

$$Af = F, \quad f \in \mathcal{F} \quad (1.23)$$

if some functions which form the equation are unknown, but data are given. Using these data the approximations to the unknown functions can be obtained. Let $F_t(\lambda)$ and $F_t(y, x)$ be approximations to the distribution functions $F(x)$ and $F(y, x)$ obtained from the data.

A difference exists between the problem of density estimation and the problems of conditional probability and conditional density estimation. In the problem of density estimation, instead of an accurate right-hand side of the

[†]Actually, the solution of this equation is the definition of conditional density. Suppose that $p(x)$ and $p(y, x)$ are the densities corresponding to probability distribution functions $F(x)$ and $F(y, x)$. Then equality (1.21) is equivalent to the equality $p(y|x)p(x) = p(y, x)$.

equation we have its approximation. We would like to get an approximation to the solution of Eq. (1.23) from the relationship

$$Af \approx F_\ell, \quad f \in \mathcal{F}.$$

In the problems of conditional probability and conditional density estimation, not only the right-hand side of Eq. (1.23) is known approximately, but the operator A is known approximately as well (in the left-hand side of integral equations (1.20) and (1.21), instead of the distribution functions, we use their approximations). So our problem is to get an approximation to the solution of Eq. (1.23) from the relationship

$$A_\ell f \approx F_\ell, \quad f \in \mathcal{F},$$

where A_ℓ is an approximation of the operator A .

The good news about solving these problems is that the empirical distribution function forms a good approximation to the unknown distribution function. In the next section we show that as the number of observations tends to infinity, the empirical distribution function converges to the desired one. Moreover, we shall give an asymptotically exact rate of the convergence for different metrics determining different definitions of a distance between functions.

The bad news is that the problem of solving operator equation (1.23) is the so-called ill-posed problem. In Section 1.12 we shall define the concept of "ill-posed" problems and describe the difficulties that arise when one needs to solve ill-posed problems. In the appendix to this chapter we provide the classical theory of solving ill-posed problems which is generalized in Chapter 7 to the case of stochastic ill-posed problems. The theory of solving stochastic ill-posed problems will be used for solving our integral equations.

1.11 GLIVENKO-CANTELLI THEOREM

In the 1930s Glivenko and Cantelli proved one of the most important theorems in statistics. They proved that when the number of observations tends to infinity, the empirical distribution function $F_\ell(x)$ converges to the actual distribution function $F(x)$.

This theorem and its generalizations play an important part both in learning theory and in foundations of theoretical statistics. To discuss this theorem and results related to it accurately, we need to introduce some general concepts which describe the convergence of a stochastic variable.

1.11.1 Convergence in Probability and Almost Sure Convergence

Note that an empirical distribution function is a random function because it is formed on the basis of a random sample of observations. To discuss the problem of convergence of this function we need to measure distance between the empirical distribution function and the actual one. To measure the distance between two functions, different metrics are used. In this book we use three of them: the uniform metric C

$$\rho(g_1(x), g_2(x)) = \sup_t |g_1(x) - g_2(x)|,$$

$L_2(F)$ metric

$$\rho(g_1(x), g_2(x)) = \sqrt{\int (g_1(x) - g_2(x))^2 dF(x)},$$

and $L_1(F)$ metric

$$\rho(g_1(x), g_2(x)) = \int |g_1(x) - g_2(x)| dF(x).$$

In the case when we measure the distance between random functions $F_t(x)$ and some fixed function $F(x)$, random variables

$$a_t = a_t(x_1, \dots, x_t) = \rho(F(x), F_t(x))$$

are considered. Consider a sequence of random variables

$$a_1, \dots, a_t, \dots$$

We say that a sequence of random variables a_t converges to a random variable a_0 in probability if for any $\delta > 0$ the relation

$$P\{|a_t - a_0| > \delta\} \xrightarrow{t \rightarrow \infty} 0 \quad (1.24)$$

is valid.

We say also that a sequence of random variables a_n converges to the random variable a_0 almost surely (with probability 1) if for any $\delta > 0$ the relation

$$P\{\sup_{t \geq n} |a_t - a_0| > \delta\} \xrightarrow{n \rightarrow \infty} 0 \quad (1.25)$$

is valid.

It is easy to see that the convergence (1.25) implies the convergence (1.24) which is a weaker mode of convergence. Generally, the convergence (1.24) does not imply the convergence (1.25).

The following classical lemma provides conditions under which convergence in probability implies almost sure convergence (Shiryayev, 1984).

Let A_1, \dots, A_n, \dots be a sequence of events.[†] Denote by

$$A = \overline{\lim}_{n \rightarrow \infty} A_n,$$

the event that an infinite number of events from A_1, \dots, A_n, \dots have occurred.

Lemma 1.1 (Borel–Cantelli). (a) If

$$\sum_{n=1}^{\infty} P\{A_n\} < \infty,$$

then

$$P\{\overline{\lim}_{n \rightarrow \infty} A_n\} = 0.$$

(b) If

$$\sum_{n=1}^{\infty} P\{A_n\} = \infty$$

and A_1, \dots, A_n, \dots is sequence of independent events, then

$$P\{\overline{\lim}_{n \rightarrow \infty} A_n\} = 1.$$

Corollary 1. In order for a sequence of random variables a_n to converge to a random variable a_0 almost surely, it is sufficient that for any $\delta > 0$ the inequality

$$\sum_{n=1}^{\infty} P\{ |a_n - a_0| > \delta \} < \infty$$

be fulfilled

This inequality forms necessary conditions if a_n is a sequence of independent random variables.

Corollary 2. Let $\varepsilon_n, n = 1, \dots$, be a sequence of positive values such that $\varepsilon_n \rightarrow 0$ when $n \rightarrow \infty$. Then if

$$\sum_{n=1}^{\infty} P\{ |a_n - a_0| > \varepsilon_n \} < \infty,$$

the random variables a_n converge to a random variable a_0 almost surely.

[†]See Chapter 2 for definition of events.

Convergence in probability will be denoted by

$$a_t \xrightarrow[t \rightarrow \infty]{P} a_0.$$

Almost sure convergence will be denoted by

$$a_t \xrightarrow[t \rightarrow \infty]{a.s.} a_0.$$

1.11.2 Glivenko–Cantelli Theorem

Now we can formulate the Glivenko–Cantelli theorem.

Theorem 1.1 (Glivenko–Cantelli). *The convergence*

$$\sup_x |F(x) - F_t(x)| \xrightarrow[t \rightarrow \infty]{P} 0$$

takes place

In this formulation, the Glivenko–Cantelli theorem asserts the convergence in probability,[†] in the uniform metric, of the empirical distribution function $F_t(x)$ to the actual distribution function $F(x)$.

We will not prove this theorem here, which was proved originally for the one-dimensional case. This theorem and its generalization for the multi-dimensional case will be derived from the more general assertion, which we shall prove in Chapter 4.

As soon as this theorem has been proved, the problem of the rate of convergence $F_t(x)$ to $F(x)$ emerged.

1.11.3 Three Important Statistical Laws

Investigations of the rate of convergence of $F_t(x)$ to $F(x)$ for one-dimensional continuous functions $F(x)$ resulted in the establishment of several laws of statistics, in particular the following three:

1. *Kolmogorov–Smirnov Distribution.* The random variable

$$\xi_t = \sqrt{t} \sup_x |F(x) - F_t(x)|$$

has the following limiting probability distribution (Kolmogorov):

$$\lim_{t \rightarrow \infty} P\{\sqrt{t} \sup_x |F(x) - F_t(x)| < \varepsilon\} = 1 - 2 \sum_{k=1}^{\infty} (-1)^{k-1} e^{-2\varepsilon^2 k}. \quad (1.26)$$

[†] Below we will see that almost sure convergence takes place as well.

The random variables

$$\xi_\ell^+ = \sqrt{\ell} \sup_x (F(x) - F_\ell(x)),$$

$$\xi_\ell^- = \sqrt{\ell} \sup_x (F_\ell(x) - F(x))$$

have the following limiting probability distributions (Smirnov):

$$\begin{aligned} \lim_{\ell \rightarrow \infty} P \left\{ \sqrt{\ell} \sup_x (F(x) - F_\ell(x)) < \varepsilon \right\} &= 1 - e^{-2\varepsilon^2}, \\ \lim_{\ell \rightarrow \infty} P \left\{ \sqrt{\ell} \sup_x (F_\ell(x) - F(x)) < \varepsilon \right\} &= 1 - e^{-2\varepsilon^2}. \end{aligned} \quad (1.27)$$

2. *The Law of the Iterated Logarithm.* The equality

$$P \left\{ \lim_{n \rightarrow \infty} \sup_{\ell > n} \sup_x \sqrt{\frac{2\ell}{\ln \ln \ell}} |F(x) - F_\ell(x)| = 1 \right\} = 1 \quad (1.28)$$

holds true.

3. *Smirnov Distribution.* The statistic

$$\omega^2 = \ell \int (F(x) - F_\ell(x))^2 dF(x)$$

(the so-called *omega square statistic*) has the limiting distribution

$$\begin{aligned} \lim_{\ell \rightarrow \infty} P \left\{ \ell \int (F(x) - F_\ell(x))^2 dF(x) < \varepsilon \right\} \\ = 1 - \frac{2}{\pi} \sum_{k=1}^{\infty} \int_{(2k-1)\pi}^{2k\pi} \frac{\exp\left\{-\frac{\lambda^2 \varepsilon}{2}\right\}}{\sqrt{-\lambda \sin \lambda}} d\lambda. \end{aligned}$$

We shall not prove these statistical laws. For our purpose of constructing the learning theory, we need more general laws which we shall derive in Chapters 4 and 5. Now our goal is to use the laws above to estimate the bounds for distribution function $F(x)$ provided the estimate $F_\ell(x)$.

We derive these bounds from the Kolmogorov–Smirnov law (1.27). For this purpose we consider for some η ($0 < \eta < 1$) the equality

$$1 - e^{-2\varepsilon^2 \ell} = 1 - \eta$$

which we solve with respect to ε

$$\varepsilon = \sqrt{-\frac{\ln \eta}{2\ell}}.$$

Now (1.27) can be described as follows: With probability $1 - \eta$ simultaneously for all x the inequalities

$$F_\ell(x) - \sqrt{-\frac{\ln \eta}{2\ell}} \leq F(x) \leq F_\ell(x) + \sqrt{-\frac{\ln \eta}{2\ell}} \quad (1.29)$$

are valid as $\ell \rightarrow \infty$.

Similarly, the iterated logarithm law (1.28) implies that when

$$\ell \rightarrow \infty$$

simultaneously for all x , the inequalities

$$F_\ell(x) - \sqrt{\frac{\ln \ln \ell}{2\ell}} \leq F(x) \leq F_\ell(x) + \sqrt{\frac{\ln \ln \ell}{2\ell}}$$

are valid. These inequalities are tight.

To estimate the density we have to solve an integral equation where the right-hand side of the equation is unknown, but approximations which converge to the actual function are given. But even if the approximation $F_\ell(\cdot)$ tends to $F(x)$ with a high asymptotic rate of convergence, the problem of solving our integral equations is hard, since (as we will see in the next section) it is an ill-posed problem.

1.12 ILL-POSED PROBLEMS

We say that the solution of the operator equation

$$Af(t) = F(x) \quad (1.30)$$

is *stable* if a small variation in the right-hand side $F(x) \in F(x, \alpha)$ results in a small change in the solution: that is, if for any ε there exists $\delta(\varepsilon)$ such that the inequality

$$\rho_{E_1}(f(t, \alpha_1), f(t, \alpha_2)) \leq \varepsilon$$

is valid as long as inequality

$$\rho_{E_2}(F(x, \alpha_1), F(x, \alpha_2)) \leq \delta(\varepsilon)$$

holds. Here the indices E_1 and E_2 denote that the distance is defined in the metric spaces E_1 and E_2 , respectively (the operator equation (1.30) maps functions of space E_1 into functions of space E_2).

We say that the problem of solving the operator equation (1.30) is *well-posed in the Hadamard sense* if the solution of the equation

- exists,
- is unique, and
- is stable.

The problem of solving an operator equation is considered ill-posed if the solution of this equation violates at least one of the above-mentioned requirements. In this book, we consider ill-posed problems when the solution of the operator equation exists, is unique, but is not stable.

This book considers ill-posed problems defined by the Fredholm integral equation of type I:

$$\int_a^b K(t, x) f(t) dt = F(x).$$

However, all the results obtained will also be valid for equations defined by any other linear continuous operator.

Thus, consider Fredholm's integral equation of type I:

$$\int_0^1 K(t, x) f(t) dt = F(x) \quad (1.31)$$

defined by the kernel $K(t, x)$, which is continuous almost everywhere on $0 \leq t \leq 1$, $0 \leq x \leq 1$. This kernel maps the set of functions $\{f(t)\}$, continuous on $[0, 1]$, onto the set of functions $\{F(x)\}$ also continuous on $[0, 1]$.

We shall now show that the problem of solving the equation (1.31) is an ill-posed one. For this purpose we note that the continuous function $G_\nu(x)$ which is formed by means of the kernel $K(t, x)$:

$$G_\nu(x) = \int_0^1 K(t, x) \sin \nu t dt$$

possesses the property

$$G_\nu(x) \xrightarrow{\nu \rightarrow \infty} 0.$$

Consider the integral equation

$$\int_0^1 K(t, x) f^*(t) dt = F(x) + G_\nu(x).$$

Since the Fredholm equation is linear, the solution of this equation has the form

$$f^*(t) = f(t) + \sin \nu t,$$

where $f(t)$ is the solution of Eq. (1.31). For sufficiently large ν , the right-hand side of this equation differs from the right-hand side of (1.31) only by the small amount $G_\nu(x)$, while its solution differs by the amount $\sin \nu t$.

The Fredholm integral equation is the equation we shall consider in this book. Here are some examples of problems connected with a solution of this equation:

Example 1 (The Problem of Identifying Linear Dynamic Systems). It is known that dynamic properties of linear homogeneous objects

$$y(t) = Ax(t)$$

with one output are completely described by the impulse response function $f(\tau)$. The function $f(\tau)$ is the response of the system to a unit impulse $\theta(t)$ served at the system at time $\tau = 0$.

Knowing this function, one can compute the response of the system to the disturbance $x(t)$ using the formula

$$y(t) = \int_0^t x(t - \tau) f(\tau) d\tau.$$

Thus, the determination of the dynamic characteristics of a system is reduced to the determination of the weight function $f(x)$.

It is also known that for a linear homogeneous system, the Wiener–Hopf equation

$$\int_0^\infty R_{yy}(t - \tau) f(\tau) d\tau = R_{yy}(t) \quad (1.32)$$

is valid.

Equation (1.32) connects the autocorrelation function $R_{yy}(u)$ of a stationary random process at the input of the object with the weight function $f(\tau)$ and the joint correlation function of the input and output signals $R_{yy}(t)$.

Thus, the problem of identifying a linear system involves determining the weight function based on the known autocorrelation function of the input signal and the measured (observed) joint correlation function of the input and output signals; that is, it is a problem of solving integral equation (1.32) on the basis of empirical data.

Example 2 (The Problem of Estimating Derivatives). Let measurements of a smooth function $F(x)$ at ℓ points of the interval $[0,1]$ be given. Suppose that the points at which the measurements were taken are distributed randomly and independently according to the uniform distribution. The problem is to estimate the derivative $f(x)$ of the function $F(x)$ on $[0, 1]$.

It is easy to see that the problem is reduced to solving the Volterra integral equation of type I,

$$\int_0^x f(t) dt = F(x) - F(0),$$

under the condition that the ℓ measurements

of the function $F(x)$ at the points

$$x_1, \dots, x_\ell$$

are known. Equivalently, it is reduced to the solution of the Fredholm equation of the type I,

$$\int_0^1 \theta(x-t)f(t)dt = F(x) - F(0),$$

where

$$\theta(u) = \begin{cases} 1 & \text{if } u > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Note that in the case when $F(x)$ is a monotonically increasing function satisfying the conditions $F(0) = 0$, $F(1) = 1$, we have the problem of density estimation.

In the general case when the k th derivative has to be estimated, the following integral equation has to be solved:

$$\int_0^1 \frac{(x-t)^{k-1}}{(k-1)!} \theta(x-t)f(t)dt = F(x) - \sum_{j=0}^{k-1} \frac{F^{(j)}(0)}{j!},$$

where in place of $F(x)$ the empirical data y_1, \dots, y_ℓ are used. Here $F^{(j)}(0)$ is the value of the j th derivative at zero.

The main difficulty in solving integral equations stems from the fact that this is an ill-posed problem since the solution of the equation is *unstable*. In the mid-1960s, several methods for solving unstable problems of mathematical physics were proposed. In the appendix to this chapter, we shall present the so-called "regularization method" proposed by A. N. Tikhonov. This method is applicable for solving integral equations when instead of knowing the function on the right-hand side of an equation, one knows the sequence of approximations which converges to an unknown function with probability one.

In the 1970s we generalized the theory of the regularization method for solving the so-called stochastic ill-posed problems. We define stochastic ill-posed problems as problems of solving operator equations in the case when approximations of the function on the right-hand side converge in probability to an unknown function and/or when the approximations to the operator converge in probability to an unknown operator. This generalization will be presented in Chapter 7. We show that the regularization method solves stochastic ill-posed problems as well. In particular, it solves our learning problems: estimating densities, conditional densities, and conditional probabilities.

1.13 THE STRUCTURE OF THE LEARNING THEORY

Thus, in this chapter we have considered two approaches to learning problems. The first approach (imitating the supervisor's operator) brought us to the problem of minimizing a risk functional on the basis of empirical data.

The second approach (identifying the supervisor's operator) brought us to the problem of solving some integral equation when the elements of an equation are known only approximately.

It has been shown that the second approach gives more details on the solution of pattern recognition and regression estimation problems.

Why in this case do we need both approaches'? As we mentioned in the last section. the second approach, which is based on the solution of the integral equation, forms an ill-posed problem. For ill-posed problems, the best that can be done is to obtain a sequence of approximations to the solution which converges in probability to the desired function when the number of observations tends to infinity. For this approach, there exists no way to evaluate how well the problem can be solved if a finite number of observations is used. In the framework of this approach to the learning problem, any exact assertion is asymptotic.

That is why the first approach, based on minimizing the risk functional from empirical data of the finite size ℓ , may be more appropriate for our purposes.

In the following chapters, we show that in the framework of the first approach one can estimate how close the risk functional of the chosen function is to the smallest possible one (for a given set of functions).

This means that if the function $Q(z, \alpha_\ell)$ has been chosen via an appropriate induction principle (for example, the principle of empirical risk minimization), one can assert that with probability $1 - \eta$ the value of the risk $R(\alpha_\ell)$ for this function does not exceed the smallest possible value of risk $\inf_{\alpha} R(\alpha)$ (for a given set of functions) by more than ε . Here ε depends only on η , ℓ and one more parameter describing some general properties (capacity) of a given set of functions.

In other words, it will be shown that for algorithms selecting functional dependencies based on empirical risk minimization induction principles, one can guarantee that with probability at least $1 - \eta$ the inequality

$$R(\alpha_\ell) - \inf_{\alpha \in \Lambda} R(\alpha) \leq \varepsilon(\ell, \eta, \cdot) \quad (1.33)$$

holds true.

Recall that for the pattern recognition problem the goal is to obtain the solution for which the value of risk is ε -close to minimal (see Section 1.3).

For the regression estimation problem, the ε -closeness of the risk functional to the minimal one guarantees that the chosen function is $\sqrt{\varepsilon}$ -close to the regression function in the $L_2(F)$ metric (see Section 1.4).

For the density estimation problem, the ε -closeness of the risk functionals

to the minimal one implies the $(2\sqrt{1 - \exp\{-\varepsilon\}})$ -closeness of approximation to the actual density in the $L_1(F)$ metric (see Section 1.5).

Therefore the main problem in this approach (both theoretical and practical) is to find the method which provides the smallest ε on the right-hand side of inequality (1.33) (for a given number of observations).

To do this well, four levels of the theory should be developed. These are:

1. ***Theory of Consistency of the Learning Processes.*** The goal of this part of the learning theory is to give a complete description of the conceptual (asymptotic) models of the learning processes—that is, to find the necessary and sufficient conditions of consistency of the learning processes. (Informally, the conditions for convergence to zero of the ε in (1.33) as the number of observations ℓ tends to infinity. The exact definition of consistency is given in Chapter 3.)

Why do we need this asymptotic (conceptual) part of the theory if our goal is to obtain the best solution for a finite number of observations? The conceptual (asymptotic) part of the learning theory is important since to find the condition for consistency one has to introduce some concepts in terms of which the theory can be developed. For example, the concept which characterizes the capacity of a given set of functions (the dot in arguments of ε in the inequality (1.33)). Generally, it is possible to use several different constructions. However, it is important to develop the theory on the basis of such constructions which are not only sufficient for the consistency of learning process, but are *necessary* as well. This gives us a guarantee that the theory which we develop using these constructions is general and from the conceptual point of view cannot be improved.

2. ***Theory of Estimating the Rate of Convergence of the Learning Processes.*** This part of the learning theory is devoted to obtaining nonasymptotic bounds on the generalization ability of the learning machines (ε on the right-hand side of inequality (1.33)). We obtain these bounds using the concepts developed in the conceptual part of the theory. In this book, we consider a theory of distribution-free bounds of the rate of convergence (the theory that does not use a priori information about the unknown probability measure). The main requirement of this part of the theory is to find a way to construct bounds for different sets of functions.
3. ***Theory for Controlling the Rate of Convergence of the Learning Processes.*** The bounds on generalization ability will be used for developing the new induction principles that guarantee the best solution of the learning problem for a given finite set of observations.

These induction principles are based on the trade-off between complexity of the chosen function (capacity of the set of functions from which the function is chosen) and the value of empirical risk which can be achieved using this function. This trade-off led to some functional different from the empirical risk functional that should be minimized.

Table 1.1. Structure of Learning Theory and Its Representation in this Book

Parts of the Theory	Chapters	Content of the Chapters
1. Theory of consistency of the learning processes	Chapter 3 Chapter 1-1 Chapter 15 Chapter 16	Review of the theory Proofs of the theorems Proofs of the theorems Proofs of the theorems
2. Theory of bounds	Chapter 4 Chapter 5	For indicator functions For real-valued functions
3. Theory of controlling the generalization	Chapter 6 Chapter 7 Chapter 8	SRM induction principle Stochastic ill-posed problems New setting of the problem
4. Theory of the learning algorithms and its applications	Chapter 9 Chapter 10 Chapter 11 Chapter 12 Chapter 13	Classical approaches SVM for pattern recognition SVM for function estimation Examples of pattern recognition Examples of function estimation

Obtaining these functionals in explicit form is the goal of this part of the theory.

4. *Theory of the Algorithms.* Finally, there is a theory of learning algorithms. The goal of this part of the theory is to develop tools for minimizing the functionals describing the trade-off. In order to minimize these functionals, it is necessary to develop algorithms which can control both the minimization of empirical risk in a given set of functions and the choice of a set of functions with appropriate capacity.

In this book, we consider all parts of the theory of minimization of the risk functional from empirical data.

We consider the theory of solving stochastic ill-posed problem as well, and we apply it to estimate density, conditional density, and conditional probability. This theory describes sufficient conditions for consistency of the solution and, for some cases, the asymptotic rate of convergence of the solution. Of course, the results of asymptotic theory is not enough to guarantee the success if the algorithms use limited samples. In the framework of this theory, our hope is that asymptotic properties established in the theory are also valid for not very large ℓ .

Table 1.1 shows the structure of the learning theory and its representation in this book.

(Chapter 2 is not indicated in this table. The content of that chapter goes beyond the learning theory. It, however, is very important for the general understanding of the nature of learning problems. We show in this chapter how deeply these problems are connected with fundamental problems of theoretical statistics.

APPENDIX TO CHAPTER 1: METHODS FOR SOLVING ILL-POSED PROBLEMS

A1.1 THE PROBLEM OF SOLVING AN OPERATOR EQUATION

We say that two sets of elements $f \in M$ and $F \in N$ are connected by *functional dependency* if to each element $f \in M$ there corresponds a unique element $F \in N$.

This functional dependence is called a *function* if the sets M and N are sets of numbers; it is called a *functional* if M is a set of functions and N is a set of numbers, and it is called an operator if both sets are sets of functions.

Each operator A uniquely maps elements of the set M onto elements of the set N . This is denoted by the equality

$$AM = N.$$

In a collection of operators we shall single out those which realize a one-to-one mapping of M into N . For these operators the problem of solving the operator equation

$$Af(t) = F(x) \tag{A1.1}$$

can be considered as the problem of finding an element $f(t)$ in M to which an element $F(x)$ corresponds in N .

For operators which realize a one-to-one mapping of elements M onto N and a function $F(x) \in N$, there exists a unique solution of the operator equation (A.1). However, finding a method for solving an operator equation of such generality is a hopeless task. Therefore we shall investigate operator equations with continuous operators only.

Let the elements $f \in M$ belong to a metric space E_1 with metric $\rho_1(\cdot, \cdot)$, and the elements $F \in N$ belong to a metric space E_2 with metric $\rho_2(\cdot, \cdot)$. An

operator A is called *continuous* if "close" elements (with respect to metric ρ_1) in E_1 are mapped into "close" elements (with respect to metric ρ_2) in E_2 .

We shall consider an operator equation defined by a continuous one-to-one operator M onto N . The solution of such an operator equation exists and is unique, that is, there exists inverse operator A^{-1} from N onto M :

$$M = A^{-1}N$$

The basic problem is whether the inverse operator is continuous.

If the operator A^{-1} is continuous, then close preimages will correspond to close function in N , that is, the solution of the operator equation (A1.1) will be *stable*.

If, however, the inverse operator is not continuous, then the solution of the operator equation can be *nonstable*. In this case according to Hadamard's definition (Chapter 1, Section 1.12), the problem of solving an operator equation is ill-posed.

It turns out that in many important cases, for example, for a so-called completely continuous operator A , the inverse operator A^{-1} is not continuous and hence the problem of solving the corresponding operator equation is ill-posed.

Definition. We say that a linear operator A defined in a linear normed space E_1 with the range of values in a linear normed space E_2 is *completely continuous* if it maps any bounded set of the functions in the space E_1 into a compact set of the space E_2 —that is, if each bounded infinite sequence in E_1

$$f_1, f_2, \dots, f_i, \dots \quad \|f_i\| \leq c, \quad (\text{A1.2})$$

(here $\|f_i\|$ is the norm in E_1) is mapped in E_2 into a sequence

$$Af_1, \dots, Af_i, \dots \quad (\text{A1.3})$$

such that a convergent subsequence

$$Af_{t_1}, \dots, Af_{t_k}, \dots \quad (\text{A1.4})$$

can be extracted from it

We will show that if the space E_1 contains bounded noncompact sets, then the inverse operator A^{-1} for an absolutely continuous operator A need not be continuous.

Indeed, consider a bounded noncompact set in E_1 . Select in this set an infinite sequence (A1.2) such that no subsequence of it is convergent. An infinite sequence (A1.3) from which convergent subsequence (A1.4) may be

selected (since operator A is absolutely continuous) corresponds in E_2 to this sequence. If the operator A^{-1} were continuous, then a convergent sequence

$$f_{i_1}, \dots, f_{i_k}, \dots, \quad (\text{A1.5})$$

would correspond to the sequence (A1.4) in E_1 which is a subsequence of (A1.2). This, however, contradicts the choice of (A1.2).

Thus, the problem of solving an operator equation defined by a completely continuous operator is an ill-posed problem. In the main part of this book we shall consider linear integral operators

$$Af = \int_a^b K(t, x)f(t) dt$$

with the kernel $K(t, x)$ continuous in the domain $a \leq t \leq b$, $a \leq x \leq b$. These operators are completely continuous from $C[a, b]$ into $C[a, b]$. The proof of this fact can be found in textbooks on functional analysis (see, for example. Kolmogorov and Fomin (1970)).

A1.2 PROBLEMS WELL-POSED IN TIKHONOV'S SENSE

Definition. The problem of solving the operator equation

$$Af = F \quad (\text{A1.6})$$

is called *well-posed (correct) in Tikhonov's sense* on the set $M^* \subset M$, and the set M^* is called the set (class) of correctness, provided that:

1. The solution of (A1.6) exists for each $F \in AM^* = N^*$ and belongs to M^* .
2. The solution belonging to M^* is unique for any $F \in N^*$.
3. The solutions belonging to M^* are stable with respect to $F \in N^*$.

If $M^* = M$ and $N^* = N$, then correctness in Tikhonov's sense corresponds to correctness in Hadamard's sense. The meaning of Tikhonov's correctness is that correctness can be achieved by restricting the set of solutions M to a class of correctness M^* .

The following lemma shows that if we narrow the set of solutions to a compact set M^* , then it constitutes a correctness class.

Lemma. *If A is a continuous one-to-one operator defined on a compact set $M^* \subset M$, then the inverse operator A^{-1} is continuous on the set $N^* = AM'$.*

Proof. Choose an arbitrary element $F_0 \in \mathcal{N}^*$ and an arbitrary sequence convergent to it:

$$\{F_n\} \subset \mathcal{N}^*, \quad F_n \xrightarrow{n \rightarrow \infty} F_0.$$

It is required to verify the convergence

$$f_n = A^{-1}F_n \xrightarrow{n \rightarrow \infty} A^{-1}F_0 = f_0.$$

Since $\{f_n\} \subset \mathcal{M}^*$, and \mathcal{M}^* is a compact set, the limit points of the sequence $\{f_n\}$ belong to \mathcal{M}^* . Let f_0 be such a limit point. Since f_0 is a limit point, there exists a sequence $\{f_{n_k}\}$ convergent to it, to which there corresponds a sequence $\{F_{n_k}\}$ convergent to F_0 . Therefore, approaching the limit in the equality

$$Af_{n_k} = F_{n_k}$$

and utilizing the continuity of the operator A , we obtain

$$Af_0 = F_0.$$

Since the operator A^{-1} is unique, we have

$$A^{-1}F_0 = f_0$$

which implies the uniqueness of the limit point of the sequence $\{f_{n_k}\}$.

It remains to verify that the whole sequence $\{f_{n_k}\}$ converges to f_0 . Indeed, if the whole sequence is not convergent to f_0 , one could find a neighborhood of the point f_0 outside of which there would be infinitely many members of the sequence $\{f_{n_k}\}$. Since \mathcal{M}^* is compact, this sequence possesses a limit point f_0^* which, by what has been proven above, coincides with f_0 . This, however, contradicts the assumption that the selected sequence lies outside a neighborhood of point f_0 .

The lemma is thus proved.

Hence correctness in Tikhonov's sense on a compactum \mathcal{M}^* follows from the conditions of the existence and uniqueness of a solution of an operator equation. The third condition (the stability of the solution) is automatically satisfied. This fact is essentially the basis for all constructive ideas for solving ill-posed problems. We shall consider one of them.

A1.3 THE REGULARIZATION METHOD

A1.3.1 Idea of Regularization Method

The regularization method was proposed by A. N. Tikhonov in 1963.

Suppose that it is required to solve the operator equation

$$Af = F \quad (\text{A1.7})$$

defined by a continuous one-to-one operator A acting from M into N . Suppose the solution of (A1.7) exists.

Consider a lower semicontinuous functional $W(f)$, which we shall call the *regularizer* and which possesses the following three properties:

1. The solution of the operator equation belongs to the domain of definition $D(W)$ of the functional $W(f)$.
2. On the domain of the definition, functional $W(f)$ admits real-valued nonnegative values.
3. The sets

$$\mathcal{M}_c = \{ f : W(f) \leq c \}, \quad c \geq 0,$$

are all compact.

The idea of regularization is to find a solution for (A1.7) as an element minimizing a certain functional. It is not the functional

$$\rho = \rho_2(Af, F)$$

(this problem would be equivalent to the solution of Eq. (A1.7) and therefore would also be ill-posed) but is an "improved" functional

$$R_\gamma(\hat{f}, F) = \rho_2^2(A\hat{f}, F) + \gamma W(\hat{f}), \quad \hat{f} \in D(W) \quad (\text{A1.8})$$

with *regularization parameter* $\gamma > 0$. We will prove that the problem of minimizing the functional (A1.8) is stable, that is, to the close functions F and F_δ (where $\rho_2(F, F_\delta) \leq \delta$) there correspond close elements f^γ and f_δ^γ which minimize the functionals $R_\gamma(f, F)$ and $R_\gamma(f, F_\delta)$.

A1.3.2 Main Theorems About the Regularization Method

The problem in the theory of regularization is to determine a relationship between δ and γ such that the sequence of solutions f_δ^γ of regularized problems $R_\gamma(f, F_\delta)$ converges as $\delta \rightarrow 0$ to the solution of the operator equation (A1.7).

The following theorem establishes these relations.

Theorem 1. *Let E_1 and E_2 be metric spaces, and suppose for $F \in N$ there exists a solution $f \in D(W)$ of Eq. (A1.7). Let instead of an exact right-hand*

side F of Eq. (A1.7), approximations[†] $F_\delta \in E_2$ be given such that $\rho_2(F, F_\delta) \leq \delta$. Suppose the values of parameter γ are chosen in such a manner that

$$\begin{aligned} \gamma(\delta) &\rightarrow 0 \quad \text{for } \delta \rightarrow 0, \\ \lim_{\delta \rightarrow 0} \frac{\delta^2}{\gamma(\delta)} &< r < \infty. \end{aligned} \quad (\text{A1.9})$$

Then the elements $f_\delta^{\gamma(\delta)}$ minimizing the functionals $R_{\gamma(\delta)}(f, F_\delta)$ on $D(W)$ converge to the exact solution f as $\delta \rightarrow 0$.

Proof. The proof of the theorem utilizes the following fact: For any fixed $y > 0$ and an arbitrary $F \in \mathcal{N}$ an element $f^y \in D(W)$ exists which minimizes the functional $R_y(f, F)$ on $D(W)$.

Let y and δ satisfy the relation (A1.9). Consider a sequence of elements $f_\delta^{\gamma(\delta)}$ minimizing $R_{\gamma(\delta)}(f, F_\delta)$, and show that the convergence

$$f_\delta^{\gamma(\delta)} \xrightarrow[\delta \rightarrow 0]{} f$$

is valid.

Indeed, by definition of $f_\delta^{\gamma(\delta)}$ we have

$$\begin{aligned} R_{\gamma(\delta)}(f_\delta^{\gamma(\delta)}, F_\delta) &\leq R_{\gamma(\delta)}(f, F_\delta) = \rho_2^2(Af, F_\delta) + \gamma(\delta)W(f) \\ &\leq \delta^2 + \gamma(\delta)W(f) = \gamma(\delta) \left(W(f) + \frac{\delta^2}{\gamma(\delta)} \right). \end{aligned}$$

Taking into account that

$$R_{\gamma(\delta)}(f_\delta^{\gamma(\delta)}, F_\delta) = \rho_2^2(Af_\delta^{\gamma(\delta)}, F_\delta) + \gamma(\delta)W(f_\delta^{\gamma(\delta)})$$

we conclude

$$W(f_\delta^{\gamma(\delta)}) \leq W(f) + \frac{\delta^2}{\gamma(\delta)}.$$

$$\rho_2^2(Af_\delta^{\gamma(\delta)}, F_\delta) \leq \gamma(\delta) \left(W(f) + \frac{\delta^2}{\gamma(\delta)} \right)$$

Since the conditions (A1.9) are fulfilled, all the elements of the sequence $f_\delta^{\gamma(\delta)}$ for a $\delta > 0$ sufficiently small belong to a compactum \mathcal{M}_c , where $c^* = W(f) + r + \varepsilon > 0$, $\varepsilon > 0$, and their images $F_\delta^{\gamma(\delta)} - Af_\delta^{\gamma(\delta)}$ are convergent:

$$\begin{aligned} \rho_2(F_\delta^{\gamma(\delta)}, F) &\leq \rho_2(F_\delta^{\gamma(\delta)}, F_\delta) + \delta \\ &\leq \delta + \sqrt{\delta^2 + \gamma(\delta)W(f)} \xrightarrow[\delta \rightarrow 0]{} 0. \end{aligned}$$

[†]The elements F_δ need not belong to the set \mathcal{N} .

This implies, in view of the lemma, that their preimages

$$f_\delta^{\gamma(\delta)} \rightarrow f \quad \text{for } \delta \rightarrow 0$$

are also converged.

The theorem is thus proved.

In a Hilbert space the functional $W(f)$ may be chosen to be equal to $\|f\|^2$ for a linear operator A. Although the sets \mathcal{M}_c are (only) weakly compact in this case, the convergence of regularized solutions—in view of the properties of Hilbert spaces—will be, as shown below, a strong one. Such a choice of a regularizing functional is convenient also because its domain of definition $D(W)$ coincides with the whole space E_1 . However, in this case the conditions imposed on the parameter γ are more rigid than in the case of Theorem 1; namely, γ should converge to zero slower than δ^2 .

Thus the following theorem is valid.

Theorem 2. *Let E_1 be a Hilbert space and $W(f) = \|f\|^2$. Then for $\gamma(\delta)$ satisfying the relations (A1.9) with $r = 0$, the regularized elements $f_\delta^{\gamma(\delta)}$ converge as $\delta \rightarrow 0$ to the exact solution f in the metric of the space E_1 .*

Proof. It is known from the geometry of Hilbert spaces that the sphere $\|f\| \leq c$ is a weak compactum and that from the properties of weak convergence of elements f_i to the element f and convergence of the norms $\|f_i\|$ to $\|f\|$ there follows the strong convergence

$$\|f_i - f\| \xrightarrow[i \rightarrow \infty]{} 0.$$

Moreover, it follows from the weak convergence $f_i \rightarrow f$ that

$$\|f\| \leq \liminf_{i \rightarrow \infty} \|f_i\| \tag{A1.10}$$

Utilizing these properties of Hilbert spaces, we shall now prove the theorem.

It is not difficult to check that for a weak convergence in the space E_1 the preceding theorem is valid: $f_\delta^{\gamma(\delta)}$ converges weakly to f as $\delta \rightarrow 0$. Therefore in view of (A1.10) the inequality

$$\|f\| \leq \liminf_{\delta \rightarrow 0} \|f_\delta^{\gamma(\delta)}\|$$

is valid. On the other hand, taking into account that $W(f) = \|f\|^2$ and that $r = 0$, we obtain

$$\limsup_{\delta \rightarrow 0} \|f_\delta^{\gamma(\delta)}\|^2 \leq \lim_{\delta \rightarrow 0} \left(\|f\|^2 + \frac{\delta^2}{\gamma(\delta)} \right) = \|f\|^2.$$

Hence the convergence of the norms is valid:

$$\|f_\delta^{\gamma(\delta)}\| \xrightarrow[\delta \rightarrow 0]{} \|f\|,$$

and along with it the validity of weak convergence implies, in view of the properties of Hilbert spaces, the strong convergence

$$\|f_\delta^{\gamma(\delta)} - f\| \xrightarrow[\delta \rightarrow 0]{} 0.$$

The theorem is thus proved.

The theorems presented above are fundamentals in regularization theory. Using these theorems the feasibility of solving ill-posed problems is established.

In Chapter 7 we consider the so-called stochastic ill-posed problems and generalize these theorems for the stochastic case. Using the method of regularization for stochastic ill-posed problems we consider our learning problems of estimating densities, conditional probabilities, and conditional densities.

2

ESTIMATION OF THE PROBABILITY MEASURE AND PROBLEM OF LEARNING

The two approaches to the learning problem presented in the first chapter were not chosen accidentally. These approaches correspond to two different cases for which the estimation of probability measure on the basis of empirical data is possible. Recall that the common part in both approaches is the fact that the probability measure (distribution function) is unknown and the information about the measure has to be extracted from the data.

Generally, however, it is impossible to estimate a probability measure using only empirical data.

One can estimate the probability measure if:

1. The measure belongs to specific sets of measures or
2. The measure is estimated partially.

These two options for estimating probability measures imply two different approaches to the statement of the learning problem.

2.1 PROBABILITY MODEL OF A RANDOM EXPERIMENT

The goal of this chapter is to demonstrate that the analysis of consistency of the learning processes is in many ways equivalent to the analysis of the core problem of statistics—estimation of probability measure.

To start the discussion about different ways of estimating probability measure based on the results of a random experiment, let us briefly recall the model of a random experiment used in probability theory. This model is described in advanced textbooks on probability theory.

According to Kolmogorov's axiomatization, to every random experiment there corresponds a set Ω of elementary events w which defines all the possible outcomes of an experiment (the elementary events). On the set Ω of elementary events, a system $\{\mathbf{A}\}$ of subsets $\mathbf{A} \in \Omega$, which are called *events*, is defined. The entire set Ω considered as an event determines a situation corresponding to the sure event (an event that always occurs). It is assumed that the set $\{\mathbf{A}\}$ contains the empty set \emptyset , describing the event that never occurs.

For the set $\{\mathbf{A}\}$ the following operations are defined: *union*, *complement*, and *intersection*.

On the set Ω the σ -algebra \mathcal{F} of the events \mathbf{A} is defined. The set \mathcal{F} of subsets of Ω is called the σ -algebra of events $\mathbf{A} \in \Omega$ if the following hold:

1. $\Omega \in \mathcal{F}$.
2. If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$.
3. If $A_i \in \mathcal{F}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

The pair (Ω, \mathcal{F}) is an idealization of the *qualitative* aspect of a random experiment.

The *quantitative* aspect of an experiment is determined by the *probability measure* $P(A)$ defined on the elements \mathbf{A} of the set \mathcal{F} .

The function $P(A)$ defined on the elements $\mathbf{A} \in \mathcal{F}$ is called the *countably additive probability measure* on \mathcal{F} or, for simplicity, the *probability measure* provided that

1. $P(A) \geq 0$;
2. $P(\Omega) = 1$;
3. $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$ if $A_i, A_j \in \mathcal{F}$, and $A_i \cap A_j = \emptyset$, $i \neq j$.

We say that a probabilistic model of an experiment is specified if the probability space defined by the triple (Ω, \mathcal{F}, P) is given.

Now consider the experiment consisting of ℓ distinct trials in the probability space (Ω, \mathcal{F}, P) and let

$$w_1, \dots, w_\ell$$

be the outcomes of these trials. We say that sequence w_1, \dots, w_ℓ is a sequence of ℓ *independent* trials if for any $A_1, \dots, A_{k_\ell} \in \mathcal{F}$ the equality

$$P\{w_1 \in A_1; \dots; w_\ell \in A_{k_\ell}\} = \prod_{i=1}^{\ell} P\{w_i \in A_{k_i}\}$$

is valid.

The concept of a random variable plays an important role in stochastic analysis. Let the mapping

$$\Omega \rightarrow R^1$$

be given, performed by a real-valued function

$$\xi(w) = \xi. \quad (2.1)$$

For this random variable (function) to be measurable, we need that the relation

$$\{w : \xi(w) < z\} \in \mathcal{B} \quad (2.2)$$

be valid for any z .

Suppose that the σ -algebra \mathcal{B} of events A is related to the function $(w) = \xi$ in such a way that for any z , relation (2.2) holds true. In this case there exists the function

$$F_\xi(z) = P\{w : \xi(w) < z\} \quad (2.3)$$

which we call the *probability distribution function of the random variable ξ* .

A random vector

$$\bar{\xi}(w) = \bar{\xi}$$

determined by the mapping

$$\Omega \rightarrow R^n$$

is a generalization of the notion of a random variable.

For the vector function $\bar{\xi}(w)$ to be measurable, we need that the relation

$$\{\omega : \bar{\xi}(\omega) < \bar{z}\} \in \mathcal{F} \quad (2.4)$$

be valid for any vector \bar{z} . The inequality in the braces should be interpreted coordinatewise.

Suppose that the σ -algebra \mathcal{B} of events A is related to the function $\bar{\xi}(w) = \bar{\xi}$ in such a way that for any \bar{z} the relation (2.4) holds true. In this case there exists the function

$$F_\xi(z) = P\{w : \bar{\xi}(w) < \bar{z}\} \quad (2.5)$$

which we call the *probability distribution function of the random vector ξ* .

2.2 THE BASIC PROBLEM OF STATISTICS

2.2.1 The Basic Problems of Probability and Statistics

In the preceding section we defined a model of a random experiments by the triple (Ω, \mathcal{B}, P) . Now let

$$w_1, \dots, w_\ell$$

be the result of ℓ independent trials with the model (Ω, \mathcal{F}, P) . Consider the random variable $v_\ell(A; w_1, \dots, w_\ell)$ defined for a fixed event $A \in \mathcal{F}$ by the value

$$v_\ell(A) = v(A; w_1, \dots, w_\ell) = \frac{n_A}{\ell},$$

where n_A is the number of elements in the set w_1, \dots, w_ℓ belonging to event A . The random variable $v_\ell(A)$ is called the *frequency of occurrence of an event A* in a series of independent, random trials.

In terms of the probability distribution function of the random variable $v_\ell(A; w_1, \dots, w_\ell)$, we can formulate the basic problem of the probability theory.

Basic Problem of Probability Theory. Given model (Ω, \mathcal{F}, P) and the event $A \in \mathcal{F}$, estimate the distribution function

$$F(z; A, \ell) = P\{v_\ell(A) < z\}$$

(or some of its properties).

In this book we are concerned with the *inverse* problem. Let a qualitative model of a random experiment (Ω, \mathcal{F}) be given. Our goal is to estimate the probability measure from a given sample

$$w_1, \dots, w_\ell; \quad (2.6)$$

this means that we are attempting to estimate the values $P(A)$ for all events $A \in \mathcal{F}$.

This problem forms the basic problem of mathematical statistics.

Basic Problem of Mathematical Statistics. Given pair (Ω, \mathcal{F}) and the data (2.6) obtained from a series of random and independent trials under probability measure P , estimate this probability measure $P = \{P(A) : A \in \mathcal{F}\}$ (defined on the subsets $A \in \mathcal{F}$).

To estimate the probability measure we have to define an *estimator* $\mathcal{E}_\ell(A)$ which approximates the probability measure for all elements A of the σ -algebra of events \mathcal{F} . We want to find an estimator that defines a sequence of approximations converging to the unknown probability measure—in some modes—when the number of observations increases. To begin to analyze the possibilities of estimating the probability measure, we have to define these modes.

2.2.2 Uniform Convergence of Probability Measure Estimates

Definition. We say that the estimator

$$\mathcal{E}_\ell(A) = \mathcal{E}(A; w_1, \dots, w_\ell), \quad A \in \mathcal{F}$$

defines a sequence of measure approximations that converge *uniformly* to the probability measure \mathbf{P} if the relation

$$\sup_{A \in \mathcal{F}} |\mathbf{P}(A) - \mathcal{E}_\ell(A)| \xrightarrow[\ell \rightarrow \infty]{} 0 \quad (2.7)$$

holds true.

According to the definition of convergence in probability, this means that for any given positive ε the convergence

$$\mathbf{P}\{\sup_{A \in \mathcal{F}} |P(A) - \mathcal{E}_\ell(A)| > \varepsilon\} \xrightarrow[\ell \rightarrow \infty]{} 0 \quad (2.8)$$

takes place.

When the a-algebra \mathcal{F} of the events \mathbf{A} is poor (for example, it contains a finite number of elements \mathbf{A}), the estimator which provides uniform convergence to the probability measure can be found easily. For example, we can estimate the probability measure with frequencies at all elements \mathbf{A} of the finite σ -algebra \mathcal{F} :

$$\nu_\ell(A) = \nu(A; w_1, \dots, w_\ell) = \frac{n(A)}{\ell},$$

where $n(A)$ is the number of elements from the set w_1, \dots, w_ℓ belonging to $A \in \mathcal{F}$. The estimator $\nu_\ell(A)$ is called the *empirical measure*.

Indeed, when the number of elements A_k of a-algebra is finite and is equal to N , the following inequality is true:

$$P\left\{\sup_{1 \leq k \leq N} |P(A_k) - \nu_\ell(A_k)| > \varepsilon\right\} \leq \sum_{k=1}^N P\{|P(A_k) - \nu_\ell(A_k)| > \varepsilon\} \xrightarrow[\ell \rightarrow \infty]{} 0.$$

The convergence to zero here is due to the law of large numbers. According to this law, any summand tends to zero as the number of observations increases. Since the number of summands is finite, the whole sum converges to zero as well.

The problem of whether or not convergence (2.8) takes place arises when a-algebra \mathcal{F} is rich. For our goal of estimating the function on the basis of empirical data, we consider R^n as a space of elementary events where a-algebra \mathcal{B} is defined by Borel's sets[†]; that is, we consider the qualitative model (R^n, \mathcal{B}) .

[†]The m-algebra of Borel sets \mathcal{B} in R^1 is the smallest σ -algebra that contains all semiclosed intervals $(a, b]$.

Let \mathcal{K} be a set of parallelepipeds:

$$\Pi_{i=1}^n [a_i, b_i] = \{\xi : \xi = (\xi_1, \dots, \xi_n) : a_i \leq \xi_i < b_i, i = 1, \dots, n\}.$$

The m-algebra \mathcal{F} of Borel sets in R^n is the smallest m-algebra which contains \mathcal{K} .

For this model, the knowledge of the distribution function

$$F(z) = P\{\xi < z\}$$

is equivalent to the knowledge of the probability measure P . Thus the model in which we are interested may be described by triple (R'', \mathcal{F}, P) .

Unfortunately, for this model of random events, one cannot estimate the probability measure using the empirical measures $\nu_\ell(\mathbf{A})$.

Example. Let R be the interval $(0,1)$ and let the unknown measure given by the uniform probability distribution function be

$$F(z) = P\{\xi < z\} = z, \quad 0 < z < 1$$

Let \mathcal{F} be the Borel σ -algebra. It contains all unions of a finite number of subintervals of the interval $(0,1)$. Clearly (see Fig. 2.1), for any sample

$$\xi_1, \dots, \xi_\ell$$

and for any $\varepsilon > 0$ one can find an event $\mathbf{A}' \in \mathcal{F}$ such that two equalities

$$\nu_\ell(\mathbf{A}') = \nu(A^*; \xi_1, \dots, \xi_\ell) = 1$$

$$P(A^*) < \varepsilon$$

take place.

Thus, in this case, for any ℓ the equality

$$P\left\{\sup_{A \in \mathcal{F}} |P(A) - \nu_\ell(A)| = 1\right\} = 1 \quad (2.9)$$

holds true.

This example shows that in general the empirical estimator of probability measures does not provide the uniform convergence to the desired probability measure.



FIGURE 2.1. For any sample one can find an event A' with small probability measure ε that contains this sample.

2.3 CONDITIONS FOR THE UNIFORM CONVERGENCE OF ESTIMATES TO THE UNKNOWN PROBABILITY MEASURE

2.3.1 Structure of Distribution Function

Thus, the special estimator (empirical probability measure $\nu_\ell(A)$) does not provide the uniform convergence to any probability measure. The important question is:

Does another estimator $\mathcal{E}_\ell(A)$ exist that can provide the uniform convergence to any probability measure?

The answer is no. In general, no estimator provides uniform convergence to any unknown probability measure. To explain why this is true we need to recall some facts from the theory of probability measures.

We start with Lebesgue's theorem about the structure of probability distribution functions on the line (Shiryayev, 1984).

Theorem 2.1 (Lebesgue). *Any probability distribution function on the line can uniquely be represented as the sum*

$$F(x) = F_D(x) + F_{AC}(x) + F_S(x)$$

of three nonnegative monotone functions where:

1. F_D is a discrete component representable in the form

$$F_D(x) = \sum_{x_i < x} p(x_i), \quad p(x_i) \geq 0, \quad \sum_i p(x_i) \leq 1;$$

2. $F_{AC}(x)$ is an absolutely continuous component representable in the form

$$F_{AC}(x) = \int_{-\infty}^x p(x') dx',$$

where $p(x) \geq 0$;

3. $F_C(x)$ is a singular component—a continuous function whose set of jumps (points x for which $F(x+\varepsilon) - F(x-\varepsilon) > 0$, $\varepsilon \rightarrow 0$) has Lebesgue measure equal to zero.[†]

This theorem actually asserts that any measure on the line is a composition of three different types of measures:

[†]The standard example of a singular component is the Cantor function.

1. A measure of the first type is concentrated on at most countably many points each of which has a positive measure.
2. A measure of the second type possesses a density.
3. A measure of the third type is concentrated on a subset of the line with measure zero which has no point with positive measure.

Note that according to the Glivenko–Cantelli theorem the empirical distribution function $F_\ell(x)$ converges to the actual $F(x)$ in C metric as ℓ increases. However, to construct an estimator of probability measure which provides uniform convergence, it is necessary to find a way for estimating function $dF(x)$ rather than $F(x)$.

The following theorem asserts that in general (if an unknown probability measure contains all three components) no estimator of probability measure provides uniform convergence to the desired one.

Theorem 2.2 (Chentsov). *Let \mathcal{P}_0 be the set of all admissible probability measures on \mathcal{B} . Then for any estimator $\mathcal{E}_\ell(A)$ of an unknown probability measure defined on the Borel subsets $A \subset (0,1)$ there exists a measure $P \in \mathcal{P}_0$ for which $\mathcal{E}_\ell(A)$ does not provide uniform convergence.*

The proof of this theorem is based on the fact that there is no method that allows us to distinguish between absolutely continuous $F_{AC}(z)$ and singular distribution laws $F_S(x)$ using samples of increasing size.

This theorem implies that an estimator providing uniform convergence to the unknown probability measure can be constructed only for some special families of probability measures that do not include simultaneously both an absolutely continuous and a singular components.

Consider a special set of probability measures whose distribution functions have no singular component.

Let $\mathcal{P}_{D&AC}$ be the collection of probability measures on (K', \mathcal{B}) that have only an absolutely continuous component $F_{AC}(z)$ and a purely discrete component $F_D(\cdot)$.

'Theorem 2.3 (Chentsov). *There exists an estimator $\mathcal{E}_\ell(A)$ which provides uniform convergence to any measure in the set $\mathcal{P}_{D&AC}$.*

The proof of this theorem is based on the idea that it is possible to arrange ℓ observations in a ordered array and use each group of coinciding observations to estimate the probability of the corresponding atom in a discrete distribution function. From the remaining part of the samples we construct the empirical distribution function and smooth it as described below. Therefore, it is possible to consider approximations both to the discrete component of a distribution function and to the absolutely continuous component of the distribution function.

To estimate the discrete component, one has to note that for any given ϵ there exist a finite number $N(\epsilon)$ of points of discrete components with probability measure of at least $1 - \epsilon$. As was shown above, for a finite number of events, uniform convergence of the estimates to the probability measure takes place. For any given ϵ , one estimates the probability measure for these points and assign zero for the rest.

Therefore, the estimates of the discrete component of probability measure converge to the actual discrete component in the uniform mode.

According to the Glivenko–Cantelli theorem when the number of observations increases, the empirical distribution function converges in probability to the original absolutely continuous component of the probability distribution function. In Chapter 7 we introduce an estimator of density which in this case converges in probability to the desired one in the L_1 metric. As will be shown in Scheffe's theorem at the end of this section, in this case there exists an estimator of probability measure which converges uniformly to the desired measure.

Thus, Chentsov theorem asserts the existence of an estimator of probability measure that provides uniform convergence to the unknown measure from the set $\mathcal{P}_{D\&AC}$. However, this theorem gives no answer to the question whether this estimator is stable.

Recall that in the first chapter we called the solution stable if small variations in the information (given data) caused small variations in the results.

Chentsov theorem asserts that there exists a solution to the problem of uniform convergence to any probability measure from $\mathcal{P}_{D\&AC}$. However, this solution is unstable. Indeed, for the problem of estimating the measure to be solved, it is necessary to estimate two components of the distribution function (the discrete component and the absolutely continuous component) separately.

To estimate the discrete component, one needs to verify the exact coincidence of two observations. Such a requirement does not provide a stable method for separation of the two components of distribution functions. Thus, the methods of estimating the probability measure that contains both discrete and absolutely continuous components are unstable.

Finally, consider the set of measures \mathcal{P}_{AC} which contains only absolutely continuous functions $F_{AC}(z)$. In other words, consider measures that have densities; that is, consider measures that have the following structure

$$\int_{-\infty}^z p(x) dx = F(x). \quad (2.10)$$

The problem of density estimation on the basis of empirical data (the solution of the integral equation (2.10) when instead of the function $F(z)$ the approximation $F_t(A)$ is given) was considered in Chapter 1 as one of two approaches to the statement of the learning problem.

In spite of the fact that solving this equation is an ill-posed problem, in Chapter 7 we show that the regularization method described in Appendix to Chapter 1 provides a solution to the density estimation problem.

2.3.2 Estimator that Provides Uniform Convergence

On the basis of the estimated density $p_t(z)$, one can construct the following estimator of the probability measure:

$$\mathcal{E}_t(A) = \int_A p_t(x) dx, \quad A \in \mathcal{F}. \quad (2.11)$$

The next theorem actually shows that if the sequence of densities $p_t(z)$ converges in L_1 metric to the original one, then the estimators (2.11) of probability measure provide uniform convergence to the desired probability measure.

Let $p(x)$ and $q(x)$ be densities, let \mathcal{F} be Borel sets of events A , and let

$$P(A) = \int_A p(x) dx, \quad Q(A) = \int_A q(x) dx$$

be probabilities of the set $A \in \mathcal{F}$ corresponding to these densities. The following theorem is then valid.

Theorem 2.4 (Scheffe)

$$\sup_{A \in \mathcal{F}} |P(A) - Q(A)| = 1/2 \int |p(x) - q(x)| dx.$$

As a consequence of the Scheffe theorem, the estimator

$$\mathcal{E}_t(A) = \int_A p_t(x) dx$$

provides uniform convergence to the probability measure

$$P(A) = \int_A p(x) dx, \quad A \in \mathcal{F},$$

if the density $p_t(z)$ converges in probability to $p(z)$ in the L_1 metric

$$\int |p(x) - p_t(x)| dx \xrightarrow[t \rightarrow \infty]{P} 0. \quad (2.12)$$

Thus, uniform convergence based on estimator (2.11) to any probability measure is possible for a set of measures that can be described by the ab-

solutely continuous distribution function $F_{AC}(x)$. To construct this estimator, one has to estimate a probability density from data.

Note that this is the problem we face in the second approach considered in Chapter 1. In this approach we have to estimate densities (conditional densities) on the basis of data.

Thus, the second approach to the learning problem (identification of the supervisor's operator) is connected with uniform convergence of an estimate to the unknown probability measure.

In the next section we consider another mode of convergence: the so-called partial uniform convergence. We show that the first approach to the learning problem is based on this mode convergence to the probability measure.

2.4 PARTIAL UNIFORM CONVERGENCE AND GENERALIZATION OF GLIVENKO-CANTELLI THEOREM

2.4.1 Definition of Partial Uniform Convergence

Definition. We say that estimator $\mathcal{E}_\ell(A)$ provides partial uniform convergence to the probability measure P determined by the set of events \mathcal{F}^* if the following convergence in probability

$$\sup_{A \in \mathcal{F}^*} |P(A) - \mathcal{E}_\ell(A)| \xrightarrow[\ell \rightarrow \infty]{P} 0 \quad (2.13)$$

holds true, where $\mathcal{F}^* \subset \mathcal{F}$ is a subset of the set \mathcal{F} .

According to the definition of convergence in probability, this means that for any given positive ε the convergence

$$P \left\{ \sup_{A \in \mathcal{F}^*} |P(A) - \mathcal{E}_\ell(A)| > \varepsilon \right\} \xrightarrow[\ell \rightarrow \infty]{} 0 \quad (2.14)$$

takes place.

The difference in the definitions of uniform convergence and partial uniform convergence is in the set of events that should be taken into account when the probability measure is estimated:

1. For the uniform convergence, the supremum is taken over all elements A of the σ algebra of events \mathcal{F} .
2. For the partial uniform convergence, the supremum is taken only over the subset $\mathcal{F}^* \subset \mathcal{F}$. (The subset \mathcal{F}^* need not be a σ algebra.)

It is possible that partial uniform convergence takes place when uniform convergence fails.

Now consider random experiments, which are described by the triple (R^1, \mathbf{S}, P) . Suppose that on the space of elementary events defined by the Borel set \mathbf{S} we would like to estimate the probability measure on the basis of independent identically distributed samples:

$$\xi_1, \dots, \xi_t.$$

According to Chentsov theorem it is impossible to find an estimator that provides uniform convergence to any given probability measure.

Now consider a subset \mathcal{F}' of the set \mathcal{B} containing the elements

$$\mathcal{F}' = \{A_x : (-\infty, x), x \in (-\infty, \infty)\}. \quad (2.15)$$

In other words, \mathcal{F}' contains all sets $(-\infty, x)$.

Consider the estimator that we called the empirical measure:

$$\nu(A_x; \xi_1, \dots, \xi_t) = \frac{1}{t} \sum_{i=1}^t \theta(x - \xi_i),$$

where $\theta(u)$ is the step function:

$$\theta(u) = \begin{cases} 1 & \text{if } u > 0, \\ 0 & \text{otherwise.} \end{cases}$$

This estimator determines the frequency of any given event $A_x = (-\infty, x)$ using the examples ξ_1, \dots, ξ_t . We will use this empirical measures to estimate the unknown measure partially.

In order to show that for set \mathcal{B}' there exists a uniform convergence of the empirical measures to desired one, we need to show that for any positive ε the following relation takes place:

$$P\left\{\sup_{A \in \mathcal{F}'} |P(A) - \nu_t(A)| > \varepsilon\right\} \xrightarrow[t \rightarrow \infty]{} 0. \quad (2.16)$$

To show this, we note that by definition

$$P(A_x) = P\{\xi < x\} = F(x),$$

$$\nu_t(A_x) = F_t(x),$$

where $F(x)$ is the distribution function and $F_t(x)$ is the empirical distribution function, and (2.16) is the assertion of the Glivenko-Cantelli theorem (see Chapter 1, Section 1.10).

2.4.2 Generalization of the Glivenko–Cantelli Problem

Let us reformulate the Glivenko–Cantelli theorem in the following way:

Theorem 2.5 (Glivenko–Cantelli). *For any given probability measure $P \in \mathcal{P}_0$ and any given $\varepsilon > 0$ the following relation always hold true:*

$$P \left\{ \sup_{A_x \in \mathcal{F}'} |P(A_x) - \nu_\ell(A_x)| > \varepsilon \right\} \xrightarrow{\ell \rightarrow \infty} 0. \quad (2.17)$$

Remark. A stronger assertion follows from the Kolmogorov–Smirnov law (Chapter 1, Eq. (1.26)): The asymptotic rate of convergence does not depend on the probability measure and has an exponential form (Fig. 2.2):

$$\sup_{P \in \mathcal{P}_0} P \left\{ \sup_{A_x \in \mathcal{F}'} |P(A_x) - \nu_\ell(A_x)| > \varepsilon \right\} < 2 \exp\{-2\varepsilon^2\ell\}.$$

This formulation of the Glivenko–Cantelli theorem is extremely important because it leads to a statement of the general problem of partial uniform convergence of the probability measure estimates. Consider once more the probability

$$P \left\{ \sup_{A \in \mathcal{F}'} |P(A) - \nu_\ell(A)| > \varepsilon \right\}. \quad (2.18)$$

As follows from the Chentsov theorem, if the set of events is rich (\mathcal{F}' is a Borel set), then no estimator can provide uniform convergence to any probability measure. In particular, this is true for empirical measure estimator $\nu_\ell(A)$.

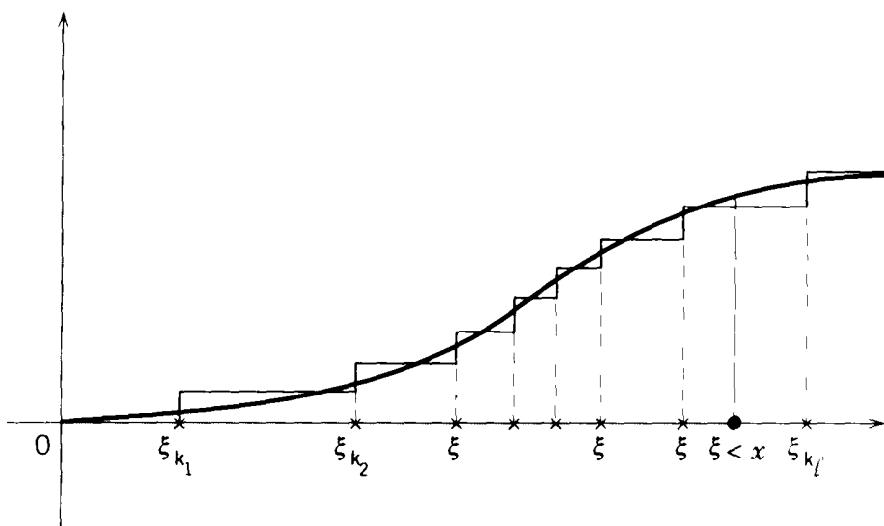


FIGURE 2.2. The uniform convergence of the frequencies to their probabilities over set of events A_x defined by all rays on the interval $x \in (-\infty, \infty)$ is equivalent to assertion of the Glivenko–Cantelli theorem.

But on the other hand, according to the Glivenko–Cantelli theorem, if the set of events is rather poor ($\mathcal{F}' = \mathcal{F}^*$), then (2.18) converges to zero for any probability measure.

The question is, What conditions should hold for the set of events \mathcal{F}' (how rich can the set of events be) to guarantee that the empirical estimator $\nu_t(A)$ provides uniform convergence with an asymptotic exponential rate of convergence which is independent on the probability measure? In other words, when do there exist for any positive ε positive constants a and b such that for sufficiently large t the inequality

$$\sup_{P \in P_0} P \left\{ \sup_{A \in \mathcal{F}'} |P(A) - \nu_t(A)| > \varepsilon \right\} < b \exp\{-ae^2t\}$$

holds true?

The problem of finding these conditions on set \mathcal{F}' can be called the *Generalized Glivenko–Cantelli problem*. As we will show in the next chapter, the solution of the Generalized Glivenko–Cantelli problem forms one of the main conceptual parts of learning theory. In this book we shall give the complete solution to this problem (we shall give necessary and sufficient conditions for existence of this inequality).

2.5 MINIMIZING THE RISK FUNCTIONAL UNDER THE CONDITION OF UNIFORM CONVERGENCE OF PROBABILITY MEASURE ESTIMATES

In Chapter 1 we considered two approaches to the learning problem. The first approach was based on the idea of minimizing the risk functional

$$R(\alpha) = \int Q(z, \alpha) dF(z) \quad (2.19)$$

in the set of functions $Q(z, \alpha)$, $\alpha \in \Lambda$, when the probability distribution function $F(z)$ is unknown but the data

$$z_1, \dots, z_n$$

are given. The second approach was based on the idea of estimating densities (conditional density, conditional probability) by solving integral equations of the type

$$\int_{-\infty}^z p(x) dx = F(z),$$

when the right-hand side of this equation is unknown but the data are given.

In Section 2.3 we showed that if one knows a priori that the distribution function is absolutely continuous then the solution of this equation on the basis of empirical data implies the solution of the problem of estimating the probability measure in the uniform mode. From this point of view, the learning problem based on the second approach is connected with the problem of

estimating the probability measure in the uniform mode, when the unknown distribution function is absolutely continuous.

Note that under conditions of uniform convergence of probability measure estimates, the functional (2.19) can be minimized as well. In this section we show that when uniform convergence takes place, one can achieve a more general solution of the learning problem than was considered in Chapter 1.

Consider the problem of minimizing the functional (2.19) on the basis of data. (The problem of minimizing (2.19) when the distribution function $F(z)$ is unknown but a random, independent sample obtained in accordance with $F(z)$ is given.)

For the time being, assume that the absolute values of the loss function $Q(z, a), a \in A$, are uniformly bounded by a quantity B . (This is always true for the pattern recognition problem.)

Let $F(z)$ be an absolutely continuous function. Then the risk functional (2.19) can be rewritten in the form

$$R(\alpha) = \int Q(z, \alpha) dF(z) = \int Q(z, \alpha)p(z) dz,$$

where $p(z)$ is the corresponding probability density.

Let us use the data to estimate the probability density $p(x)$. Assume that an estimator $p_\ell(z)$ converges in L_1 to the density $p(z)$. Consider the functional

$$R_{\text{emp}}^*(\alpha) = \int Q(z, \alpha)p_\ell(z) dz \quad (2.20)$$

defined by the means of the estimator $p_\ell(z)$.

We state the following inductive principle for minimizing risk (2.19):

As an approximation to the function $Q(z, \alpha_0)$ which yields the minimum (2.19), we shall select the function $Q(z, \alpha_\ell)$ which minimizes (2.20).

We will show that if the estimator $p_\ell(z)$ converges to $p(x)$ in L_1 , then the principle of minimizing the risk (2.20) provides solutions with risks that converge to the smallest possible risk for any set of bounded functions $Q(z, \alpha), \alpha \in \Lambda$.

Indeed, for the set of bounded functions $|Q(z, \alpha)| \leq B, a \in A$, it follows from (2.12) that

$$\begin{aligned} & \sup_{\alpha \in \Lambda} \left| \int Q(z, \alpha) dF(z) - \int Q(z, \alpha)p_\ell(z) dz \right| \\ & \leq \sup_{\alpha \in \Lambda} \int |Q(z, \alpha)| |p(z) - p_\ell(z)| dz \\ & \leq B \int |p(z) - p_\ell(z)| dz \xrightarrow[\ell \rightarrow \infty]{P} 0. \end{aligned}$$

From this relationship we derive that for any ε and any η there exists a value $\ell(\varepsilon, \eta)$ such that for any $\ell > \ell(\varepsilon, \eta)$ with probability $1 - \eta$ the following two inequalities hold true:

$$\begin{aligned} \int Q(z, \alpha_\ell) p(z) dz - \int Q(z, \alpha_\ell) p_\ell(z) dz &< \varepsilon, \\ - \int Q(z, \alpha_0) p(z) dz + \int Q(z, \alpha_0) p_\ell(z) dz &< \varepsilon. \end{aligned}$$

On the other hand, by definition,

$$\int Q(z, \alpha_\ell) p_\ell(z) dz \leq \int Q(z, \alpha_0) p_\ell(z) dz$$

Combining these three inequalities we obtain the result that with a probability of at least $1 - \eta$ the inequality

$$\int Q(z, \alpha_\ell) p(z) dz - \int Q(z, \alpha_0) p(z) dz < 2\varepsilon$$

holds true.

That means convergence in probability of the functionals

$$\int Q(z, \alpha_\ell) dF(z) \xrightarrow[\ell \rightarrow \infty]{P} \int Q(z, \alpha_0) dF(z);$$

that is, the functions $Q(z, \alpha_\ell)$ minimizing the functional (2.20) form a sequence of risks that converges in probability to the minimal one as the number of observations increases.

Thus, under the condition of uniform convergence of estimates $\mathcal{E}_\ell(A)$ to the probability measure, the induction principle (2.20) guarantees the existence of a solution which makes the risk ε -close to the minimal possible *for any uniformly bounded set of functions $Q(z, a), a \in A$* .

In the next section we show that under conditions of partial uniform convergence of probability measure estimates, the principle of empirical risk minimization provides solutions with risks that converge to the smallest possible risk (as the number of examples increase) if the set of events \mathcal{F}^* that determines the partial uniform convergence is connected with a set of functions $Q(z, a), a \in A$, in the special way.

2.6 MINIMIZING THE RISK FUNCTIONAL UNDER THE CONDITION OF PARTIAL UNIFORM CONVERGENCE OF PROBABILITY MEASURE ESTIMATES

Let us start this section by writing the Lebesgue–Stieltjes integral for the bounded nonnegative function $0 \leq Q(z, a^*) \leq B$ (here a^* is fixed) in explicit

form. According to the definition of the Lebesgue–Stieltjes integral, the functional of risk $R(\alpha^*)$ is

$$R(\alpha^*) = \int Q(z, \alpha^*) dF(z) = \lim_{n \rightarrow \infty} \sum_{i=1}^n \frac{B}{n} P \left\{ Q(z, \alpha^*) > \frac{iB}{n} \right\},$$

where

$$P \left\{ Q(z, \alpha) > \frac{iB}{n} \right\}$$

is the probability of event

$$A_i = \left\{ z : Q(z, \alpha^*) > \frac{iB}{n} \right\}$$

(see Fig. 2.3). Consider in a similar form the means of this event estimated from the data z_1, \dots, z_ℓ , $\ell = 1, \dots$

$$R_{\text{emp}}(\alpha^*) = \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha^*) = \lim_{n \rightarrow \infty} \sum_{i=1}^n \frac{B}{n} v_\ell \left\{ Q(z, \alpha^*) > \frac{iB}{n} \right\},$$

where

$$v_\ell \left\{ Q(z, \alpha^*) > \frac{iB}{n} \right\}$$

is the frequency of events A_i estimated from this data.

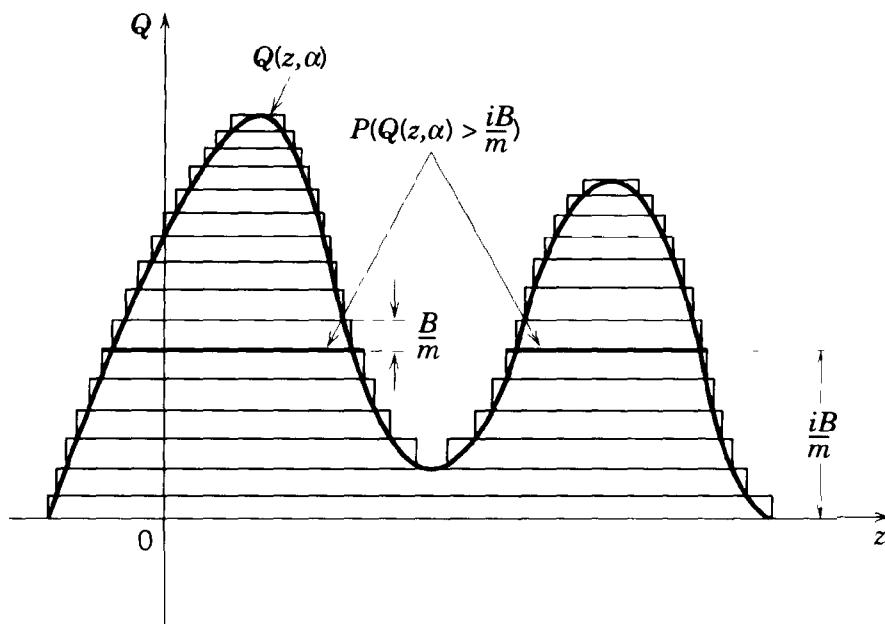


FIGURE 2.3. The Lebesgue integral of a nonnegative bounded function $0 \leq Q(z, \alpha^*) \leq B$ is the limit of a sum of products, where factor $P\{Q(z, \alpha^*) > iB/n\}$ is the (probability) measure of the set $\{z: Q(z, \alpha^*) > iB/n\}$ and the factor B/m is the height of a slice.

Now let $Q(z, \mathbf{a}), \mathbf{a} \in \mathbf{A}$ be a set of bounded functions:

$$0 \leq Q(z, \alpha) \leq B, \quad \alpha \in \Lambda.$$

Consider the following set \mathcal{F}^* of events

$$A_{\alpha, \beta} = \{\mathbf{z}: Q(z, \alpha) \geq \beta\}, \quad \alpha \in \Lambda, \beta \in [0, B].$$

Suppose that the empirical estimator $v_\ell(\mathbf{A})$ defines measures that partially converge to the unknown probability measure P :

$$\sup_{A \in \mathcal{F}^*} |P(A) - v_\ell(A)| \xrightarrow[\ell \rightarrow \infty]{} 0. \quad (2.21)$$

Then, from the definitions of the Lebesgue–Stieltjes integral and of the partial uniform convergence (2.21) we obtain

$$\begin{aligned} & \sup_{\alpha \in \Lambda} \left| \int Q(z, \alpha) dF(z) - \frac{1}{l} \sum_{i=1}^l Q(z_i, \alpha) \right| \\ &= \lim_{n \rightarrow \infty} \sup_{\alpha \in \Lambda} \left| \sum_{i=1}^n \frac{B}{n} (P(A_{\alpha, iB/n}) - v_\ell(A_{\alpha, iB/n})) \right| \\ &\leq B \sup_{\alpha, \beta} |P(A_{\alpha, \beta}) - v_\ell(A_{\alpha, \beta})| \\ &= B \sup_{A \in \mathcal{F}^*} |P(A) - v_\ell(A)| \xrightarrow[\ell \rightarrow \infty]{} 0. \end{aligned} \quad (2.22)$$

It follows from this that the uniform convergence of means to their mathematical expectations on the set of uniformly bounded functions $Q(z, \mathbf{a}), \mathbf{a} \in \mathbf{A}$ is valid.

Now we prove that under the conditions of the existence of uniform convergence (2.22), the principle of minimizing the empirical risk provides a sequence of functions that converges in probability to the best solution.

As above, it follows from the uniform convergence (2.22) that for any ε and any η a value $\ell(\varepsilon, \eta)$ exists such that for any $\ell > \ell(\varepsilon, \eta)$ with probability $1 - \eta$ the following two inequalities hold true:

$$\int Q(z, \alpha_\ell) dF(z) - \frac{1}{\ell} \sum_{i=1}^\ell Q(z_i, \alpha_\ell) < \varepsilon,$$

$$- \int Q(z, \alpha_0) dF(z) + \frac{1}{\ell} \sum_{i=1}^\ell Q(z_i, \alpha_0) < \varepsilon.$$

Note that by the definition the inequality

$$\frac{1}{l} \sum_{i=1}^l Q(z_i, \alpha_\ell) \leq \frac{1}{l} \sum_{i=1}^l Q(z_i, \alpha_0)$$

is valid.

Combining these three inequalities, we obtain that with probability of at least $1 - \eta$ the inequality

$$\int Q(z, \alpha_\ell) dF(z) - \int Q(z, \alpha_0) dF(z) < 2\epsilon \quad (2.23)$$

holds true.

In other words we get convergence in probability

$$\int Q(z, \alpha_\ell) dF(z) \xrightarrow[\ell \rightarrow \infty]{P} \int Q(z, \alpha_0) dF(z); \quad (2.24)$$

that is, as the sample size increases, the functions that minimize empirical functional on $Q(z, a), a \in A$, form a sequence of values $R(\alpha_\ell)$ that converges in probability to the minimal value of the risk $R(\alpha_0)$.

2.7 REMARKS ABOUT MODES OF CONVERGENCE OF THE PROBABILITY MEASURE ESTIMATES AND STATEMENTS OF THE LEARNING PROBLEM

Earlier in this chapter we considered the basic problem of mathematical statistics: the problem of estimating a probability measure from empirical data. We showed that, in general, it is impossible to construct a universal estimator of probability measure (applicable to any probability measure).

This fact splits the study of the problem of estimating probability measure into two parts:

1. The study of conditions on probability measures $P \in \mathcal{P}$ under which the uniform convergence

$$\sup_{A \in \mathcal{F}} |P(A) - \mathcal{E}_\ell(A)| \xrightarrow[\ell \rightarrow \infty]{P} 0$$

holds true for any set \mathcal{F} .

2. The study of the conditions on sets \mathcal{F}^* under which the uniform convergence

$$\sup_{A \in \mathcal{F}^*} |P(A) - \nu_\ell(A)| \xrightarrow[\ell \rightarrow \infty]{P} 0$$

holds true for any probability measure.

The main problem in the study of uniform convergence over the entire set of σ -algebra is to define the estimator $\mathcal{E}_\ell(A)$ and the set of measures \mathcal{P} for which such convergence takes place. (In Section 2.3 it was shown that one can use the estimator (2.11).)

The main problem in the study of the partial uniform convergence is to describe conditions on the set \mathcal{F}^* under which the estimator of empirical measure provides the partial uniform convergence for any probability measure $P \in \mathcal{P}_0$ (the Generalized Glivenko–Cantelli problem).

The analysis of these two ways of estimating probability measure forms the foundation of theoretical statistics.

In the first chapter we formulated two approaches to the learning problem: One was based on the idea of imitating the supervisor's operator, while the other was based on the idea of identifying the supervisor's operator.

From the mathematical point of view the idea of identification of supervisor's operator is based on estimating the probability measure uniformly over the entire set of σ -algebra.

The idea of imitating the supervisor's operator can be described by the scheme of minimizing the risk functional on the basis of empirical data. The solution of this problem is based on partial estimating of the probability measure.

Therefore from the conceptual point of view, analysis of consistency of the learning processes is in many ways equivalent to analysis of the problem of estimating the probability measure, which is the central problem of theoretical statistics.

The next chapter is devoted to the theory of consistency of learning processes for the scheme of imitation of the supervisor's operator, while Chapter 7 is devoted to the theory of consistency for the scheme of identification of the supervisor's operator. The results obtained in these chapters can also be described in terms of convergence (in two different modes) of estimates of probability measures.

3

CONDITIONS FOR CONSISTENCY OF EMPIRICAL RISK MINIMIZATION PRINCIPLE

In this chapter we present necessary and sufficient conditions for consistency of the empirical risk minimization principle. First we formulate and prove the key theorem of the empirical risk minimization theory—the theorem about equivalence. According to this theorem, the following two facts are equivalent:

1. The principle of empirical risk minimization is consistent.
2. The specific empirical process converges.

Then we describe the theorems about the convergence of this empirical process (proofs of the theorems are the subject of the third part of this book). We show that proofs of these theorems are based on the idea of the non-falsifiability of the learning machine, where the concept of nonfalsifiability is closely related to Popper's nonfalsifiability concept introduced in philosophy of science. At the end of the chapter we discuss the necessity of the ways in which learning theory is constructed in this book.

3.1 CLASSICAL DEFINITION OF CONSISTENCY

In Chapter 1 we introduced the problem of minimizing the risk-functional

$$R(\alpha) = \int Q(z, \alpha) dF(z), \quad \alpha \in \Lambda \quad (3.1)$$

on the set of functions $Q(z, \alpha)$, $\alpha \in A$, where the distribution function $F(z)$ is unknown; however, independent identically distributed data according to

this function

$$z_1, \dots, z_\ell, \quad (3.2)$$

are given.

To solve this problem, the principle of empirical risk minimization was proposed. According to this principle, instead of minimizing functional (3.1), one has to minimize the empirical risk functional

$$R_{\text{emp}}(\alpha) = \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha), \quad \alpha \in \Lambda. \quad (3.3)$$

Let

$$Q(z, \alpha_\ell) = Q(z, \alpha(z_1, \dots, z_\ell))$$

be a function that minimizes[†] the functional (3.3). The fundamental problem of empirical risk minimization theory is to describe the situations where this principle is consistent. Below we give a classical definition of consistency.

Definition. We say that the principle (method) of empirical risk minimization is *consistent* for the set of functions $Q(z, a), a \in A$, and for the probability distribution function $F(z)$ if the following two sequences converge in probability to the same limit:

$$R(\alpha_\ell) \xrightarrow[\ell \rightarrow \infty]{P} \inf_{\alpha \in \Lambda} R(\alpha) \quad (3.4)$$

$$R_{\text{emp}}(\alpha_\ell) \xrightarrow[\ell \rightarrow \infty]{P} \inf_{\alpha \in \Lambda} R(\alpha). \quad (3.5)$$

In other words the empirical risk minimization method is consistent if it provides the sequence of functions $Q(z, \alpha_\ell), \ell = 1, 2, \dots$, for which both the expected risk and the empirical risk converge in probability to the minimal possible (for a given set of functions) value of risk (Fig 3.1). Equation (3.4) asserts that the sequence of values of achieved risks converges to the smallest possible risk for the given set of functions, and Eq. (3.5) asserts that the limit of a sequence of empirical risks estimates the minimal possible value of the risk.

The goal of this chapter is to describe conditions for consistency of the empirical risk minimization method. We would like to obtain these conditions in terms of general characteristics of a set of functions and probability measure. Unfortunately, for the classical definition of consistency given above, this is impossible since the definition includes trivial cases of consistency.

What is a trivial case of consistency?

[†]For simplicity we assume that the minimum of empirical risk functional does exist; otherwise we choose a function that provides the value of empirical risk close to infimum.

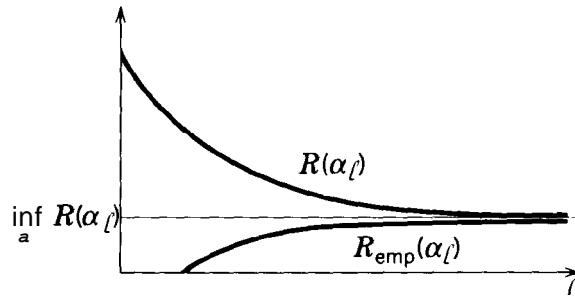


FIGURE 3.1. The learning process is consistent if both the expected risks $R(\alpha_l)$ and the empirical risks $R_{\text{emp}}(\alpha_l)$ converge to the minimal possible value of the risk, $\inf_{\alpha \in \Lambda} R(\alpha)$.

Suppose we have established that for some set of functions $Q(z, a), a \in A$, the method of empirical risk minimization is not consistent. Consider the extended set of functions which includes the initial set of functions and one additional function, $\phi(z)$. Suppose that the additional function satisfies the inequality

$$\inf_{\alpha \in \Lambda} Q(z, a) > \phi(z).$$

It is clear (Fig. 3.2) that for the extended set of functions [containing $\phi(z)$] the method of empirical risk minimization is consistent. Indeed, for any distribution function and any number of observations the minimum of the empirical risk is attained at the function $\phi(z)$ that gives the minimum of the expected risk.

This example shows that there exist trivial cases of consistency that depend on whether a given set of functions contains a minorizing function.

Therefore, any theory of consistency that uses the classical definition needs

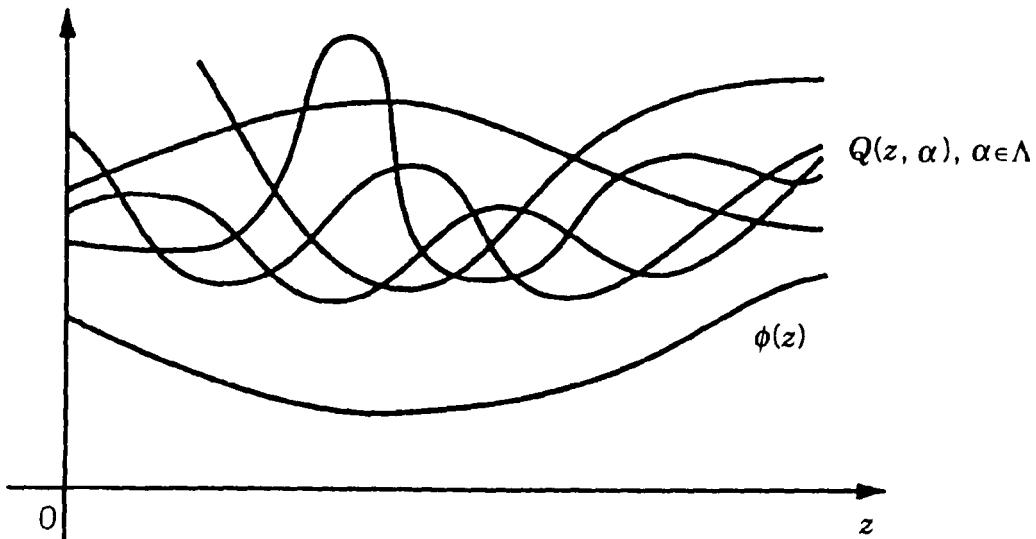


FIGURE 3.2. A case of trivial consistency. The ERM method is inconsistent on the set of functions $Q(z, \alpha), \alpha \in A$, and is consistent on the set of functions $\phi(z) \cup Q(z, \alpha), \alpha \in \Lambda$.

to check whether the case of trivial consistency is possible. That means that the theory should take into account specific functions in a given set. Our goal, however, is to find conditions for consistency that could be easily checked. We would like to get the conditions that depend on general properties of a set of functions and do not depend on specific functions in a set.

3.2 DEFINITION OF STRICT (NONTRIVIAL) CONSISTENCY

In order to develop the theory of consistency of the empirical risk minimization method which does not depend on the properties of elements of a set of functions, but depends only on general properties (capacity) of this set of functions, we need to adjust the definition of consistency to exclude the trivial consistency case.

3.2.1 Definition of Strict Consistency for the Pattern Recognition and the Regression Estimation Problems

Definition. We say that the method of minimizing empirical risk is strictly (nontrivially) consistent for the set of functions $Q(z, a)$, $a \in A$, and the probability distribution function $F(z)$ if for any nonempty subset $\Lambda(c)$, $c \in (-\infty, \infty)$, of this set of functions such that

$$\Lambda(c) = \left\{ \alpha : \int Q(z, a) dF(z) \geq c \right\}$$

the convergence

$$\inf_{\alpha \in \Lambda(c)} R_{\text{emp}}(\alpha) \xrightarrow[\ell \rightarrow \infty]{P} \inf_{\alpha \in \Lambda(c)} R(\alpha) \quad (3.6)$$

is valid.

In other words, the method of empirical risk minimization is strictly consistent if it provides convergence (3.6) both for the given set of functions and for all subsets $\Lambda(c)$ of functions that remain after the functions with the values of the risks smaller than c are excluded from this set.

Note that according to the classical definition of consistency, described in the previous section, the method is consistent if it satisfies two conditions: (3.4) and (3.5). In the definition of strict consistency we use only one condition. The following Lemma shows that under the condition of strict consistency the other condition is satisfied automatically.

Lemma. If the method of empirical risk minimization is strictly consistent, the following convergence in probability holds

$$R(\alpha_\ell) \xrightarrow[\ell \rightarrow \infty]{P} \inf_{\alpha \in \Lambda} R(\alpha).$$

Proof: Denote

$$\inf_{\alpha \in \Lambda} R(\alpha) = \int Q(z, \alpha_0) dF(z) = T.$$

For an arbitrary $\varepsilon > 0$, consider the subset $A(T + \varepsilon)$ of the set of functions $Q(z, a)$, $a \in A$, such that

$$\Lambda(T + \varepsilon) = \left\{ \alpha : \int Q(z, \alpha) dF(z) \geq T + \varepsilon \right\}.$$

We choose ε such that $A(T + \varepsilon)$ is not empty. Let (3.6) be satisfied. Then the equalities

$$\lim_{\ell \rightarrow \infty} P \left\{ \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha_\ell) \geq T + \frac{\varepsilon}{2} \right\} = 0,$$

$$\lim_{\ell \rightarrow \infty} P \left\{ \inf_{\alpha \in \Lambda(T + \varepsilon)} \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \geq T + \frac{\varepsilon}{2} \right\} = 1$$

are valid.

These equalities imply

$$\lim_{\ell \rightarrow \infty} P \{ \alpha_\ell \in \Lambda(T + \varepsilon) \} = 0.$$

If on the other hand, $\alpha_\ell \notin \Lambda(T + \varepsilon)$, then the inequality

$$T \leq \int Q(z, \alpha_\ell) dF(z) \leq T + \varepsilon$$

holds. This inequality implies (3.4).

Thus, we have proven that strict consistency implies the convergence (3.4), but not vice versa. The following example demonstrates that in some cases the convergence (3.4) does exist and the convergence (3.5) does not.

Example. Consider the following set of indicator functions $Q(z, a)$, $a \in A$, which are defined on the interval $[0, 1]$. Each function of this set is equal to 1 for all z except a finite number of intervals of measure ε where it is equal to 0 (Fig. 3.3). The parameters a define the intervals at which the function is equal to zero. The set of functions $Q(z, a)$, $a \in A$, is such that for any finite number of points z_1, \dots, z_ℓ , one can find a function that takes the value of zero on this set of points. Let $F(z)$ be the uniform distribution function on the interval $[0, 1]$.

For this set of functions, the equalities

$$R_{\text{emp}}(\alpha_\ell) = \inf_{\alpha \in \Lambda} \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) = 0,$$

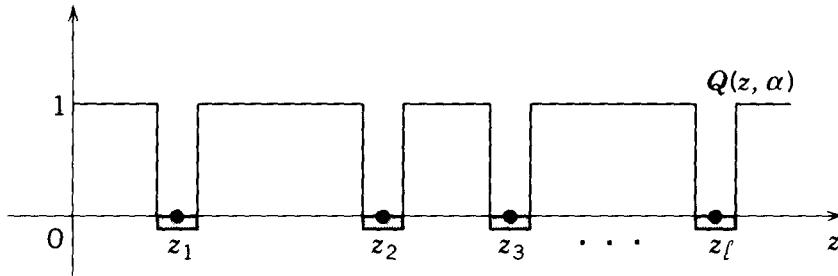


FIGURE 3.3. An example where the convergence (3.4) does exist and the convergence (3.5) does not.

$$R(\alpha) = \int Q(z, \alpha) dF(z) = 1 - \varepsilon,$$

hold. So for this set of functions, the relation

$$\inf_{\alpha \in \Lambda} R(\alpha) - R_{\text{emp}}(\alpha_\ell) = 1 - \varepsilon$$

takes place. On the other hand, for any function of this set (including the function $Q(z, \alpha_\ell)$), the relation

$$R(\alpha_\ell) - \inf_{\alpha \in \Lambda} R(\alpha) = \int Q(z, \alpha_\ell) dF(z) - \inf_{\alpha \in \Lambda} \int Q(z, \alpha) dF(z) = 0$$

holds true.

Thus, in this example, convergence (3.4) takes place though convergence (3.5) does not.

3.2.2 Definition of Strict Consistency for the Density Estimation Problem

In Chapter 1 we showed that for the density estimation problem (in the Fisher–Wald setting) the principle of empirical risk minimization implies the maximum likelihood method. For this problem the loss function associated with the set of densities $p(z, a)$, $a \in A$ (where the optimal $p(z, \alpha_0)$ is being searched for), has the form

$$Q(z, \alpha) = -\log p(z, \alpha), \quad \alpha \in \Lambda.$$

To minimize the functional

$$R(\alpha) = - \int p(z, \alpha_0) \log p(z, \alpha) dz$$

with unknown density $p(z, \alpha_0)$ using the data

we minimize the empirical risk functional

$$R_{\text{emp}}(\alpha) = -\frac{1}{\ell} \sum_{i=1}^{\ell} \log p(z_i, \alpha)$$

(the maximum likelihood method).

For the case of density estimation by the maximum likelihood method, we will use another definition of strict consistency which requires consistency for estimating any density in a given set of densities.

Definition. The maximum likelihood method is *strictly* consistent with respect to the set of densities $p(z, a), a \in A$, if for any $p(z, \alpha_0), \alpha_0 \in A$. the relation

$$\inf_{\alpha \in A} \frac{1}{\ell} \sum_{i=1}^{\ell} (-\log p(z_i, \alpha)) \xrightarrow[\ell \rightarrow \infty]{P} \int p(z, \alpha_0) (-\log p(z, \alpha_0)) dz$$

holds true where i.i.d. samples z_1, \dots, z_ℓ are drawn from the density $p(z, \alpha_0)$.

Below, we consider necessary and sufficient conditions of strict consistency both defined for the method of minimizing empirical risk and defined for the method of maximum likelihood. In this chapter we shall refer to strict consistency as simply consistency.

3.3 EMPIRICAL PROCESSES

The analysis of consistency of the empirical risk minimization method is essentially connected with the analysis of the convergence of two empirical processes.

Let the probability distribution function $F(z)$ be defined on the space $z \in R^n$, and let $Q(z, a), a \in A$, be a set of measurable (with respect to this distribution) functions. Let

$$z_1, \dots, z_\ell, \dots$$

be a sequence of independent identically distributed vectors.

Consider the sequence of random variables

$$\xi^\ell = \sup_{\alpha \in A} \left| \int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \right|, \quad \ell = 1, 2, \dots \quad (3.7)$$

We call this sequence of random variables that depends both on the probability measure $F(z)$ and on the set of functions $Q(z, a), a \in A$, a two-sided

empirical process. The problem is to describe conditions under which this empirical process converges in probability to zero.

In other words, the problem is to describe conditions such that for any positive ε the convergence

$$P \left\{ \sup_{\alpha \in A} \left| \int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \right| > \varepsilon \right\} \xrightarrow[\ell \rightarrow \infty]{} 0 \quad (3.8)$$

takes place.

We call this relation *uniform convergence of means to their mathematical expectations over a given set of functions* or, for simplicity, *uniform convergence*.

Along with the empirical process ξ^ℓ , we consider a *one-sided empirical process* given by the sequence of random values

$$\xi_+^\ell = \sup_{\alpha \in A} \left(\int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \right)_+, \quad \ell = 1, 2, \dots, \quad (3.9)$$

where we denote

$$(u)_+ = \begin{cases} u & \text{if } u > 0, \\ 0 & \text{otherwise.} \end{cases}$$

The problem is to describe conditions such that for any positive ε , the following relation

$$P \left\{ \sup_{\alpha \in A} \left(\int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \right) > \varepsilon \right\} \xrightarrow[\ell \rightarrow \infty]{} 0 \quad (3.10)$$

takes place.

We call this relation *uniform one-sided convergence of means to their mathematical expectations over a given set of functions* or, simply, *uniform one-sided convergence*.

In Chapter 2, when we considered the generalization of the Glivenko–Cantelli theorem, we actually considered special cases of the empirical process (3.8): namely, the case where the set $Q(z, a), a \in A$, was a set of indicator functions. This case will play an important part in our considerations. For a set of indicator functions, the empirical process (3.8) determines uniform convergence of frequencies to their probabilities, and process (3.10) determines uniform one-sided convergence of frequencies to their probabilities.

3.3.1 Remark on the Law of Large Numbers and Its Generalization

Note that if the set of functions $Q(z, a), a \in A$ contains only *one* element, then the sequence of random variables ξ^ℓ defined in Eq. (3.7) always con-

verges in probability to zero. This fact constitutes the main law of statistics, the Law of Large Numbers:

The sequence of means converges to expectation of a random variable (if it exists) as the number ℓ increases.

It is easy to generalize the Law of Large Numbers for the case where a set of functions $Q(z, a), a \in A$, has a finite number of elements. In contrast to the cases with a finite number of elements, the sequence of random variables ξ^ℓ for a set $Q(z, a), a \in A$, with an infinite number of elements does not necessarily converge to zero. The problem is:

To describe the properties of the set of functions $Q(z, a)$, $a \in A$, and the probability measure $F(z)$ under which the sequence of random variables ξ^ℓ converges in probability to zero.

In this case, one says that the *Law of Large Numbers in a functional space* (space of functions $Q(z, a), a \in A$) takes place or that there exists uniform (two-sided) convergence of the means to their expectation over a given set of functions.

Thus, the problem of the existence of the Law of Large Numbers in a functional space (uniform two-sided convergence of the means to their expectations) can be considered as a generalization of the classical Law of Large Numbers.

Note that in classical statistics the problem of existence of uniform one-sided convergence has not been considered; it became important due to the Key Theorem (which we formulate in the next section) pointing the way for analysis of the problem of consistency of the ERM inductive principle.

The uniform convergence (3.8) means that for sufficiently large ℓ , the empirical risk functional approximates the risk functional uniformly well over all functions in a given set of functions. In Chapter 2, Section 2.6 we showed that when uniform convergence takes place, the function which minimizes empirical risk provides the value of the risk that is close to the smallest possible risk.

So the uniform convergence gives *sufficient* conditions for the consistency of the empirical risk minimization method. In this situation arises the question:

Is it possible that the requirement of uniform convergence is too strong? Can there exist a situation such that the empirical risk minimization method is consistent, but at the same time, the uniform convergence does not take place?

In the next section we show that such a situation is *impossible*. We show that one-sided uniform convergence forms not only the sufficient condi-

tions for the consistency of the empirical risk minimization method, but the *necessary* conditions as well.[†]

3.4 THE KEY THEOREM OF LEARNING THEORY (THEOREM ABOUT EQUIVALENCE)

In this section we formulate the key theorem of learning theory which we prove in the next section. We show that for strict consistency of the empirical risk minimization method, it is *necessary and sufficient* that one-sided uniform convergence over a given set of functions takes place.

Theorem 3.1. *Let there exist the constants a and A such that for all functions in the set $Q(z, a)$, $a \in A$, and for a given distribution function $F(z)$, the inequalities*

$$a \leq \int Q(z, \alpha) dF(z) \leq A, \quad \alpha \in \Lambda$$

hold true.

Then the following two statements are equivalent:

1. *For the given distribution function $F(z)$, the empirical risk minimization method is strictly consistent on the set of functions $Q(z, a), a \in A$.*
2. *For the given distribution function $F(z)$, the uniform one-sided convergence of the means to their mathematical expectation takes place over the set of functions $Q(z, a), a \in A$.*

This theorem is stated for some fixed probability measure. However, the main interest of learning theory is to find conditions under which the empirical risk minimization method is consistent for *any* probability measure in the set \mathcal{P} . (If we have no a priori information about the problem at hand, and the distribution that performs the generator of random vectors, then the set $\mathcal{P} = \mathcal{P}_0$ is the set of all possible probability measures.) The following corollary describes conditions of consistency for the set of distribution functions:

Corollary. *Let there exist such constants a and A that for all functions in the set $Q(z, a), a \in A$, and all distribution functions: $F = F(z)$ in the set \mathcal{P} , the inequalities*

$$a \leq \int Q(z, \alpha) dF(z) \leq A, \quad \alpha \in \Lambda, \quad F(z) \in \mathcal{P}$$

hold true.

[†]Note that necessary and sufficient conditions for consistency of the learning processes is given by uniform one-sided convergence but not two-sided because we face a nonsymmetric situation: We are looking for consistency of results in **minimizing** the empirical risk, but we do not bother about consistency of results in **maximizing** the empirical risk.

Then the following two statements are equivalent:

1. For any distribution function in the set \mathcal{P} , the empirical risk minimization method is strictly consistent on the set of functions $Q(z, a), a \in A$.
2. For any distribution function in the set \mathcal{P} , the uniform one-sided convergence of the means to their mathematical expectation takes place on the set of functions $Q(z, a), a \in A$.

3.5 PROOF OF THE KEY THEOREM

Let the empirical risk minimization method be strictly consistent on the set of functions $Q(z, a), a \in A$. According to the definition of strict consistency (for a fixed measure) this means that for any c such that the set

$$\Lambda(c) = \left\{ \alpha : \int Q(z, \alpha) dF(z) \geq c \right\}$$

is nonempty the following convergence in probability is true:

$$\inf_{\alpha \in \Lambda(c)} \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \xrightarrow[\ell \rightarrow \infty]{P} \inf_{\alpha \in \Lambda(c)} \int Q(z, \alpha) dF(z). \quad (3.11)$$

Consider a finite sequence of numbers a_1, \dots, a_n , such that

$$|a_{i+1} - a_i| < \frac{\varepsilon}{2}, \quad a_1 = a, \quad a_n = A.$$

We denote by T_k the event

$$\inf_{\alpha \in \Lambda(a_k)} \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) < \inf_{\alpha \in \Lambda(a_k)} \int Q(z, \alpha) dF(z) - \frac{\varepsilon}{2}.$$

Then by virtue of (3.11),

$$P(T_k) \xrightarrow[\ell \rightarrow \infty]{} 0. \quad (3.12)$$

We denote

$$T = \bigcup_{k=1}^n T_k.$$

Since n is finite and for any k the relation (3.11) is true, it follows that

$$P(T) \xrightarrow[\ell \rightarrow \infty]{} 0. \quad (3.13)$$

We denote by A the event

$$\sup_{\alpha \in \Lambda} \left(\int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \right) > \varepsilon$$

Suppose that A takes place. Then there will be $\alpha^* \in \Lambda$ such that

$$\int Q(z, \alpha^*) dF(z) - \varepsilon > \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha^*).$$

From α^* we find k such that $\alpha^* \in \Lambda(a_k)$ and

$$\int Q(z, \alpha^*) dF(z) - a_k < \frac{\varepsilon}{2}.$$

For the chosen set $\Lambda(a_k)$ the inequality

$$\int Q(z, \alpha^*) dF(z) - \inf_{\alpha \in \Lambda(a_k)} \int Q(z, \alpha) dF(z) < \frac{\varepsilon}{2}$$

holds true.

Therefore for the chosen a' and $\Lambda(a_k)$, the following inequalities hold:

$$\begin{aligned} \inf_{\alpha \in \Lambda(a_k)} \int Q(z, \alpha) dF(z) - \frac{\varepsilon}{2} &> \int Q(z, \alpha^*) dF(z) - \varepsilon \\ &> \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha^*) \geq \inf_{\alpha \in \Lambda(a_k)} \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha), \end{aligned}$$

that is, the event T_k does occur and, hence, so does T.

Therefore,

$$P(\mathcal{A}) < P(T).$$

By (3.13),

$$\lim_{\ell \rightarrow \infty} P(T) = 0,$$

which expresses uniform one-sided convergence

$$P \left\{ \sup_{\alpha \in \Lambda} \left(\int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \right) > \varepsilon \right\} \xrightarrow[\ell \rightarrow \infty]{} 0. \quad (3.14)$$

The first part of the theorem is proved.

Now suppose that uniform one-sided convergence (3.14) takes place. Let us prove that in this case the strict consistency takes place—that is, that for any ε the convergence

$$\lim_{\ell \rightarrow \infty} P \left\{ \left| \inf_{\alpha \in \Lambda(c)} \int Q(z, \alpha) dF(z) - \inf_{\alpha \in \Lambda(c)} \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \right| > \varepsilon \right\} = 0$$

holds. Let us denote by **A** the event

$$\left| \inf_{\alpha \in \Lambda(c)} \int Q(z, \alpha) dF(z) - \inf_{\alpha \in \Lambda(c)} \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \right| > \varepsilon.$$

Then the event **A** is the union of the two events

$$\mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2$$

where

$$\mathbf{A} = \left\{ z : \inf_{\alpha \in \Lambda(c)} \int Q(z, \alpha) dF(z) + \varepsilon < \inf_{\alpha \in \Lambda(c)} \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \right\},$$

and

$$\mathcal{A}_2 = \left\{ z : \inf_{\alpha \in \Lambda(c)} \int Q(z, \alpha) dF(z) - \varepsilon > \inf_{\alpha \in \Lambda(c)} \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \right\}.$$

The goal is to bound the probability of the event **A**

$$P(\mathcal{A}) \leq P(\mathcal{A}_1) + P(\mathcal{A}_2)$$

Suppose that the event \mathcal{A}_1 occurs. To bound $P(\mathcal{A}_1)$ we take a function $Q(z, \mathbf{a}^*)$ such that

$$\int Q(z, \mathbf{a}^*) dF(z) < \inf_{\alpha \in \Lambda(c)} \int Q(z, \alpha) dF(z) + \frac{\varepsilon}{2}.$$

Then the inequality

$$\frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \mathbf{a}^*) > \int Q(z, \mathbf{a}^*) dF(z) + \frac{\varepsilon}{2}$$

holds. The probability of this inequality is therefore not less than probability of the event \mathcal{A}_1 :

$$P(\mathcal{A}_1) \leq P \left\{ \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \mathbf{a}^*) - \int Q(z, \mathbf{a}^*) dF(z) > \frac{\varepsilon}{2} \right\} \xrightarrow{\ell \rightarrow \infty} 0. \quad (3.15)$$

The probability on the right-hand side tends to zero by the law of large numbers.

If, on the other hand, the event \mathcal{A}_2 occurs, then there is a function $Q(z, \alpha^{**})$, $\alpha^{**} \in \Lambda(c)$ such that

$$\frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha^{**}) + \frac{\varepsilon}{2} < \inf_{\alpha \in \Lambda(c)} \int Q(z, \alpha) dF(z) < \int Q(z, \alpha^{**}) dF(z),$$

and, therefore,

$$\begin{aligned} P(\mathcal{A}_2) &< P \left\{ \int Q(z, \alpha^{**}) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha^{**}) > \frac{\varepsilon}{2} \right\} \\ &< P \left\{ \sup_{\alpha \in \Lambda} \left(\int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \right) > \frac{\varepsilon}{2} \right\} \xrightarrow{\ell \rightarrow \infty} 0 \end{aligned} \quad (3.16)$$

by virtue of (3.14).

Because

$$P(\mathcal{A}) \leq P(\mathcal{A}_1) + P(\mathcal{A}_2)$$

from (3.15) and (3.16) we conclude that

$$P(\mathcal{A}) \xrightarrow{\ell \rightarrow \infty} 0.$$

The theorem is proven.

3.6 STRICT CONSISTENCY OF THE MAXIMUM LIKELIHOOD METHOD

As was shown in Chapter 1, the empirical risk minimization method encompasses the maximum likelihood method. However, for the maximum likelihood method, we define another concept of strict consistency. This definition requires that for any density $p(x, \alpha_0)$ from the given set of densities $p(x, \alpha)$, $\alpha \in \Lambda$, the convergence in probability

$$\inf_{\alpha \in \Lambda} \frac{1}{\ell} \sum_{i=1}^{\ell} (-\log p(x_i, \alpha)) \xrightarrow[\ell \rightarrow \infty]{P} \inf_{\alpha \in \Lambda} \int (-\log p(x, \alpha)) p(x, \alpha_0) dx$$

has to be valid.

For the consistency of the maximum likelihood method on a set of densities the following theorem is true:

Theorem 3.2. *For the maximum likelihood method to be strictly consistent on the set of densities*

$$0 < a \leq p(x, \alpha) \leq A < \infty, \quad \alpha \in \Lambda$$

it is necessary and sufficient that uniform one-sided convergence takes place for the set of functions

$$Q(x, a) = -\log p(x, a), \quad a \in A$$

with respect to some measure $p(x, \alpha_0)$, $\alpha_0 \in A$.

Remark. It will be clear from the proof that this theorem contains in implicit form the following assertion: If one-sided uniform convergence on the set of functions

$$Q(x, \alpha) = -\log p(x, \alpha), \quad \alpha \in \Lambda$$

takes place with respect to some density $p(x, \alpha_0)$, then it will take place with respect to any density of the set $p(x, a)$, $a \in A$.

This theorem will be proved in Chapter 16.

Thus, the theorems about equivalence replaced the problem of the strict consistency of the empirical risk minimization method with the problem of existence of uniform one-sided convergence of means to their mathematical expectations or, in other words, with convergence of some empirical process. The third part of this book is devoted to studying in detail the convergence of appropriate empirical processes. However, in the next sections of this chapter we describe the main results of these studies.

3.7 NECESSARY AND SUFFICIENT CONDITIONS FOR UNIFORM CONVERGENCE OF FREQUENCIES TO THEIR PROBABILITIES

3.7.1 Three Cases of Uniform Convergence

Up until now in our consideration of the problem of risk minimization from empirical data, we did not care what specific properties the set of functions $Q(z, a)$, $a \in A$ has. Now, to describe the necessary and sufficient conditions for uniform convergence (in this section we consider the problem of uniform two-sided convergence, rather than uniform one-sided convergence), we will distinguish between three classes of functions:

1. First, we consider sets of *indicator functions* $Q(z, a)$, $a \in A$. For this set of functions, we formulate the necessary and sufficient conditions for uniform convergence of frequencies to their probabilities.

2. Then, we generalize this result for *uniformly bounded* sets of functions $Q(z, a), a \in A$. The set of functions $Q(z, a), a \in A$, is uniformly bounded if there exists a constant C such that for any function in this set, the inequality $|Q(z, a)| \leq C$ is valid. For such sets of functions, we describe the necessary and sufficient conditions for uniform convergence of means to their mathematical expectations.
3. Lastly, using the results for uniformly bounded set of functions we will describe the necessary and sufficient conditions for uniform convergence of means to their mathematical expectations for the general case, namely, when $Q(z, a), a \in A$, is a set of unbounded functions.

Thus, we shall obtain the general result in three steps.

3.7.2 Conditions of Uniform Convergence in the Simplest Model

Now let $Q(z, a), a \in A$, be a set of indicator functions. Our goal is to describe the necessary and sufficient conditions for uniform two-sided convergence—that is, the convergence

$$P \left\{ \sup_{\alpha \in \Lambda} \left| \int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \right| > \varepsilon \right\} \xrightarrow[\ell \rightarrow \infty]{} 0 \quad (3.17)$$

for any $\varepsilon > 0$.

For the set of indicator functions $Q(z, a), a \in A$, we can rewrite (3.17) in the form

$$P \left\{ \sup_{\alpha \in \Lambda} |P\{Q(z, \alpha) > 0\} - \nu_{\ell}\{Q(z, \alpha) > 0\}| > \varepsilon \right\} \xrightarrow[\ell \rightarrow \infty]{} 0, \quad (3.18)$$

where $P\{Q(z, a) > 0\}$ are probabilities of the events $A_{\alpha} = \{z : Q(z, a) > 0\}$, $a \in A$, and $\nu_{\ell}\{Q(z, a) > 0\}$ are frequencies of these events obtained on the given data z_1, \dots, z_{ℓ} .

According to the Bernoulli theorem for any fixed event $A^* = \{z : Q(z, a^*) > 0\}$, the frequencies converge to the probability when the number of observations tends to infinity. The inequality

$$P\{|P\{Q(z, a^*) > 0\} - \nu_{\ell}\{Q(z, a^*) > 0\}| > \varepsilon\} \leq 2 \exp\{-2\varepsilon^2\ell\} \quad (3.19)$$

(Chernoff inequality) describes the rate of convergence.

Our goal, however, is to describe the conditions for uniform convergence (3.17) over the set of events $A_{\alpha} = \{z : Q(z, a) > 0\}$, $a \in A$. Let us start with the simplest model.

The Simplest Model. Let our set of events contain a finite number N of events $A_k = \{z : Q(z, \alpha_k) > 0\}$, $k = 1, 2, \dots, N$. For this set of events, uniform convergence does hold. Indeed, the following sequence of inequalities is valid:

$$\begin{aligned} P\left\{\max_{1 \leq k \leq N} |P\{Q(z, \alpha_k) > 0\} - \nu_\ell\{Q(z_i, \alpha_k) > 0\}| > \varepsilon\right\} \\ \leq \sum_{k=1}^N P\{|P\{Q(z, \alpha_k) > 0\} - \nu_\ell\{Q(z_i, \alpha_k) > 0\}| > \varepsilon\} \\ \leq 2N \exp\{-2\varepsilon^2\ell\} \end{aligned} \quad (3.20)$$

$$= 2 \exp\left\{\left(\frac{\ln N}{\ell} - 2\varepsilon^2\right)\ell\right\}. \quad (3.21)$$

(To get (3.20) we use the Chernoff inequality (3.19).) The last expression suggests that in order to obtain uniform convergence for any ε , the expression

$$\frac{\ln N}{\ell} \xrightarrow[\ell \rightarrow \infty]{} 0 \quad (3.22)$$

has to be true.

Of course, for the case when our set contains a finite number N of events and the number of observations tends to infinity, this relation is true. This also proves that for any set with a finite number of events, uniform convergence takes place. However, relations of type (3.22) will be indicative for uniform convergence also in the case where the number of events in a set is infinite.

3.7.3 Entropy of a Set of Functions

The main idea of the conditions for uniform convergence that will be described below and will be proved in Chapter 14 is as follows. Even if the set of events contains infinitely many elements, only a finite number of clusters of events is distinguishable on the given sample z_1, \dots, z_ℓ . (Two events are distinguishable on a sample if there exists at least one element in the sample that belongs to one event and does not belong to the other.)

It is clear that in this case the number of clusters is not fixed and depends both on a sample and on a given set of functions. Let us denote the number of clusters by $N^A(z_1, \dots, z_\ell)$. Roughly speaking, the idea is to substitute in (3.22) the value $N^A(z_1, \dots, z_\ell)$ that depends on the sample z_1, \dots, z_ℓ and on the set of events $A = \{z : Q(z, a) > 0\}$, $a \in A$. We will show that if $N^A(z_1, \dots, z_\ell)$ increases slowly as the sample size increases (slower than any exponential function), then (3.21) converges to zero as $\ell \rightarrow \infty$, and uniform convergence takes place.

Now we determine a new concept which we will use for constructing the necessary and sufficient conditions for uniform convergence. Let a set of

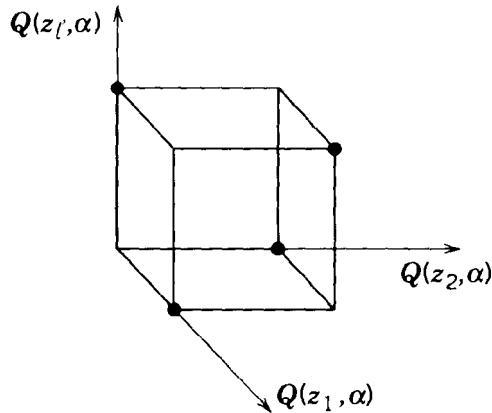


FIGURE 3.4. The set of t -dimensional binary vectors $q(\alpha)$, $\alpha \in \Lambda$, is a subset of the set of vertices of the ℓ -dimensional unit cube.

indicator functions $Q(z, \alpha)$, $\alpha \in A$ be determined on the set Z . Consider an arbitrary sequence of ℓ vectors from the set Z :

$$z_1, \dots, z_\ell. \quad (3.23)$$

Using these data, along with the set of indicator functions, let us determine the set of ℓ -dimensional binary vectors

$$q(\alpha) = (Q(z_1, \alpha), \dots, Q(z_\ell, \alpha)), \quad \alpha \in \Lambda.$$

For any fixed $\alpha = \alpha^*$ the binary vector $q(\alpha^*)$ determines some vertex of the unit cube (Fig. 3.4). Denote the number of different vertices induced both by the sample (3.23) and by the set of functions $Q(z, \alpha)$, $\alpha \in A$:

$$N^\Lambda(z_1, \dots, z_\ell).$$

It is clear that

$$N^\Lambda(z_1, \dots, z_\ell) \leq 2^\ell.$$

Let for any ℓ the function $N^\Lambda(z_1, \dots, z_\ell)$ be measurable with respect to the probability measure

$$P(z_1, \dots, z_\ell) = \prod_{i=1}^{\ell} P(z_i).$$

Definition. We say that the quantity

$$H^\Lambda(z_1, \dots, z_\ell) = \ln N^\Lambda(z_1, \dots, z_\ell)$$

is the *random entropy* of the set of indicator functions $Q(z, \alpha)$, $\alpha \in A$, on the sample z_1, \dots, z_ℓ .

We also say that the quantity

$$H^\Lambda(\ell) = \int H^\Lambda(z_1, \dots, z_\ell) dF(z_1, \dots, z_\ell)$$

is the *entropy* of the set of indicator functions $Q(z, \alpha)$, $\alpha \in \Lambda$, on samples of size ℓ .

3.7.4 Theorem About Uniform Two-Sided Convergence

Under the appropriate conditions of measurability of a set of functions the following theorem is valid.

Theorem 3.3. *In order that uniform convergence*

$$P \left\{ \sup_{\alpha \in \Lambda} \left| \int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \right| > \varepsilon \right\} \xrightarrow[\ell \rightarrow \infty]{} 0$$

over the set of indicator functions $Q(z, \alpha)$, $\alpha \in \Lambda$ be valid it is necessary and sufficient that the condition

$$\frac{H^\Lambda(\ell)}{\ell} \xrightarrow[\ell \rightarrow \infty]{} 0 \quad (3.24)$$

be satisfied.

In Chapter 14 along with Theorem 3.3 we will prove a stronger assertion:

Theorem 3.3a. *If condition (3.24) of Theorem 3.3 is satisfied, then almost sure uniform convergence takes place*

$$\sup_{\alpha \in \Lambda} \left| \int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \right| \xrightarrow[\ell \rightarrow \infty]{\text{a.s.}} 0.$$

Therefore condition (3.24) is necessary and sufficient for almost sure uniform two-sided convergence of frequencies to their probabilities.

Thus, the conditions for uniform convergence (3.24) for an infinite number of functions have the same form as for a finite number (3.22). The difference is only in characterizing the capacity of a set of functions. In the simplest case, it was the number of functions in the set; in the general case, it is the entropy of the set of indicator functions on a sample of size ℓ .

3.8 NECESSARY AND SUFFICIENT CONDITIONS FOR UNIFORM CONVERGENCE OF MEANS TO THEIR EXPECTATIONS FOR A SET OF REAL-VALUED BOUNDED FUNCTIONS

3.8.1 Entropy of a Set of Real-Valued Functions

Below, we generalize the theorem about uniform convergence obtained for sets of indicator functions to sets of real-valued functions.

We start with uniformly bounded functions $Q(z, \alpha), \alpha \in A$, where

$$|Q(z, \alpha)| < C, \quad \alpha \in A.$$

First we generalize the definition of entropy for sets of indicator functions to sets of bounded functions. As in the last section, let us consider the sequence of vectors

$$z_1, \dots, z_\ell$$

and the set of ℓ -dimensional vectors

$$q^*(\alpha) = (Q(z_1, \alpha), \dots, Q(z_\ell, \alpha)), \quad \alpha \in A.$$

The set of vectors $q^*(\alpha), \alpha \in A$, is induced both by the sample z_1, \dots, z_ℓ and by the set of uniformly bounded functions $Q(z, \alpha), \alpha \in A$.

In the last section, we considered the set of binary vectors $q(\alpha), \alpha \in A$, that was induced by the set of indicator functions $Q(z, \alpha), \alpha \in A$. For the given definition of entropy, it was important that the set of vectors $q(\alpha), \alpha \in A$ contained a finite number of elements. Now our set of vectors $q^*(\alpha), \alpha \in A$, contains an infinite number of elements, from the ℓ -dimensional cube with edges of length $2C$. Note that the set of vectors $q^*(\alpha), \alpha \in A$, belongs to the cube, but does not necessarily coincide with it (Fig. 3.5).

In mathematics, the necessity often arises to extend results valid for a finite set of elements to the infinite case. Usually such a generalization is possible if the infinite set can be covered by a *finite ϵ -net*.

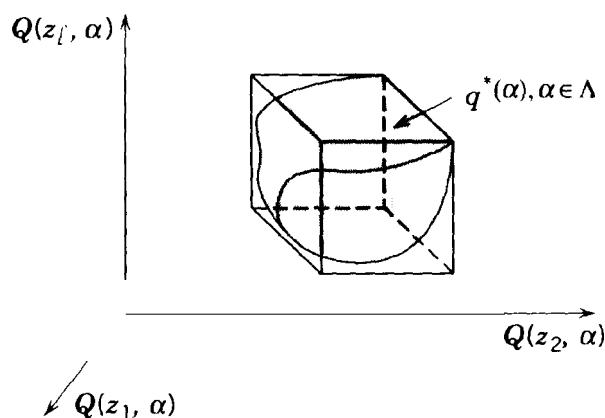


FIGURE 3.5. The set of ℓ -dimensional vectors $q(\alpha), \alpha \in A$, belongs to an ℓ -dimensional cube.

Definition. The set \mathbf{B} of elements b in a metric space M is called an ϵ -net of the set G if any point $g \in G$ is distant from some point $b \in \mathbf{B}$ by an amount not exceeding ϵ , that is,

$$\rho(b, g) < \epsilon.$$

We say that the set G admits a covering by a *finite* ϵ -net if for each ϵ there exists an ϵ -net B_ϵ consisting of a finite number of elements.

We say that the ϵ -net B_ϵ^* is a minimal ϵ -net if it is finite and contains a minimal number of elements.

To consider a minimal ϵ -net of the set of vectors $q^*(\alpha), \alpha \in A$, it is necessary to choose a metric in ℓ -dimensional Euclidean space. In Chapter 15 we show that the necessary and sufficient conditions of uniform convergence can be constructed using the C metric

$$\rho_C(q^*(\alpha_1), q^*(\alpha_2)) = \max_{1 \leq k \leq \ell} |Q(z_k, \alpha_1) - Q(z_k, \alpha_2)|.$$

Let the number of elements of a minimal ϵ -net of the set of the vectors $q^*(\alpha), \alpha \in A$, be

$$N^\Lambda(\epsilon; z_1, \dots, z_\ell).$$

This number depends on the value of ϵ , on the set of functions $Q(z, \alpha)$, $\alpha \in A$, and on the random sample z_1, \dots, z_ℓ . Using this number, we introduce the concept of random entropy, and the entropy for a given set of real-valued functions. Suppose that for any ϵ the function $\ln N^\Lambda(\epsilon; z_1, \dots, z_\ell)$ is measurable.

Definition. We say that the quantity

$$H^\Lambda(\epsilon; z_1, \dots, z_\ell) = \ln N^\Lambda(\epsilon; z_1, \dots, z_\ell)$$

is the random ϵ -entropy of the set of uniformly bounded functions $Q(z, \alpha)$, $\alpha \in A$ on the sample z_1, \dots, z_ℓ .

We say also that the quantity

$$H^\Lambda(\epsilon; \ell) = \int H^\Lambda(\epsilon; z_1, \dots, z_\ell) dF(z_1, \dots, z_\ell)$$

is the ϵ -entropy of the set of uniformly bounded functions $Q(z, \alpha)$, $\alpha \in A$, on samples of size ℓ .

3.8.2 Theorem About Uniform Two-Sided Convergence

Under the appropriate conditions of measurability of a set of functions the following theorem is valid:

Theorem 3.4. *In order that uniform convergence*

$$P \left\{ \sup_{\alpha \in A} \left| \int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \right| > \varepsilon \right\} \xrightarrow[\ell \rightarrow \infty]{} 0$$

over A set of uniformly bounded functions $Q(z, \alpha), \alpha \in A$ be valid, it is necessary and sufficient that for any $\varepsilon > 0$ the conditions

$$\frac{H^A(\varepsilon; \ell)}{\ell} \xrightarrow[\ell \rightarrow \infty]{} 0 \quad (3.25)$$

be satisfied.

In Chapter 15 along with this theorem we will prove a stronger assertion:

Theorem 3.4a. *If condition (3.25) of Theorem 3.4 is satisfied, then almost sure uniform convergence takes place*

$$\sup_{\alpha \in A} \left| \int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \right| \xrightarrow[\ell \rightarrow \infty]{\text{a.s.}} 0.$$

Therefore condition (3.25) is necessary and sufficient for almost sure uniform two-sided convergence of means to their expectations.

These theorems are generalizations of the theorems for the sets of indicator functions described in the last section. Indeed, for the C metric if $\varepsilon < 1$, the number of elements of the minimal ε -net of the set of indicator functions coincides with the number of different vertices on the unit cube induced by the set of indicator functions.

3.9 NECESSARY AND SUFFICIENT CONDITIONS FOR UNIFORM CONVERGENCE OF MEANS TO THEIR EXPECTATIONS FOR SETS OF UNBOUNDED FUNCTIONS

In order to complete the description of the theory of uniform two-sided convergence, it remains to establish the necessary and sufficient conditions of uniform two-sided convergence for the general case, namely, when $Q(z, a)$, $a \in A$, is a set of arbitrary real-valued functions with bounded expectations

$$-\infty < a \leq \int Q(z, a) dF(z) \leq A < \infty, \quad Q(z, a), a \in A.$$

To state the necessary and sufficient conditions for this case we shall consider a new notion: the envelope of a set of functions.

Definition. We say that function $K(z)$ is an *envelope* of the set of functions $Q(z, \alpha), \alpha \in A$, under the probability measure $F(z)$ if

$$\sup_{\alpha \in \Lambda} |Q(z, \alpha)| \leq K(z)$$

and

$$\int K(z) dF(z) < \infty.$$

Consider along with the set of functions $Q(z, \alpha), \alpha \in A$, the set of C -bounded functions

$$Q_C(z, \alpha) = \begin{cases} C & \text{for } Q(z, \alpha) > C \\ Q(z, \alpha) & \text{for } |Q(z, \alpha)| \leq C \\ -C & \text{for } Q(z, \alpha) < -C \end{cases}$$

for $C > 0$. For any given C , the conditions of uniform convergence for the set $Q_C(z, \alpha), \alpha \in A$, are given in Theorem 3.4. The next theorem asserts that for uniform convergence on a set of arbitrary real-valued functions it is necessary and sufficient that the envelope exists and that for any C the entropy of the set of functions $Q_C(z, \alpha), \alpha \in A$, satisfies the conditions of Theorem 3.4.

Theorem 35 (Gine and Zinn). *In order that on the set of functions $Q(z, \alpha), \alpha \in A$, with bounded expectations almost sure uniform convergence*

$$\sup_{\alpha \in \Lambda} \left| \int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \right| \xrightarrow[\ell \rightarrow \infty]{\text{a.s.}} 0,$$

takes place, it is necessary and sufficient that the set of functions $Q(z, \alpha), \alpha \in A$, has the envelope $K(z)$ and that for any C and any $\epsilon > 0$ on the set of bounded functions $Q_C(z, \alpha), \alpha \in A$, conditions (3.25) be satisfied.

3.9.1 Proof of Theorem 3.5

Proof of the Sufficiency. Let there exist an envelope for the set of functions $Q(z, \alpha), \alpha \in A$:

$$\sup_{\alpha \in \Lambda} |Q(z, \alpha)| \leq K(z),$$

$$\int K(z) dF(z) < \infty.$$

and suppose that for any C and for any ϵ the set of functions $Q_C(z, \alpha), \alpha \in A$, has an entropy satisfying the condition

$$\underline{\underline{H^A(\epsilon; \ell)}} \longrightarrow 0.$$

To prove that in this case almost sure convergence

$$\sup_{\alpha \in \Lambda} \left| \int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \right| \xrightarrow[\ell \rightarrow \infty]{\text{a.s.}} 0$$

takes place we choose C such that for a given ε the inequality

$$\int (K(z) - C)_+ dF(z) \leq \varepsilon$$

is valid, where

$$(u)_+ = \max(u, 0).$$

Let $Q(z, \alpha^*)$ be a function on which the supremum is achieved. Then the inequalities

$$\begin{aligned} & \sup_{\alpha \in \Lambda} \left| \int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \right| \\ & \leq \left| \int Q_C(z, \alpha^*) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q_C(z_i, \alpha^*) \right| \\ & \quad + \frac{1}{\ell} \sum_{i=1}^{\ell} (K(z_i) - C)_+ + \int (K(z) - C)_+ dF(z) \\ & \leq \sup_{\alpha \in \Lambda} \left| \int Q_C(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q_C(z_i, \alpha) \right| + \frac{1}{\ell} \sum_{i=1}^{\ell} (K(z_i) - C)_+ + \varepsilon. \end{aligned}$$

hold. Since the first term on the right-hand side of the inequalities converges almost surely to zero (according to Theorem 3.4a), and the second term converges almost surely to a nonnegative value that is less than ε (according to the Strong Law of Large Numbers), one can assert that the whole expression on the right-hand side converges almost surely to zero. Therefore the nonnegative expression on the left-hand side converges almost surely to zero.

Proof of Necessity. Suppose almost sure uniform convergence takes place. We have to prove that:

1. There exists an envelope

$$\sup_{\alpha \in \Lambda} |Q(z, \alpha)| \leq K(z),$$

$$\int K(z) dF(z) < \infty.$$

2. For the set of functions $Q_C(z, \alpha), \alpha \in A$, defined by any $C > 0$ the equality

$$\lim_{\ell \rightarrow \infty} \frac{H^\Lambda(\varepsilon, \ell)}{\ell} = 0$$

holds true.

Existence of the envelope. We prove existence of the envelope in four steps.

1. First we prove that almost sure convergence to zero of the sequence of random variables

$$\sup_{\alpha \in \Lambda} \left| \int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \right| \rightarrow 0 \quad \text{as } \ell \rightarrow \infty$$

implies almost sure convergence to zero of the following sequence of random variables

$$\sup_{\alpha \in \Lambda} \left| \frac{\int Q(z, \alpha) dF(z) - Q(z_\ell, \alpha)}{\ell} \right| \rightarrow 0 \quad \text{as } \ell \rightarrow \infty.$$

Let us denote by $Q(z, \alpha_\ell^*)$ the function that maximizes the difference

$$Q(z_\ell, \alpha_\ell^*) = \arg \sup_{\alpha \in \Lambda} \left| \int Q(z, \alpha) dF(z) - Q(z_\ell, \alpha) \right|.$$

The following inequalities hold true:

$$\begin{aligned} & \sup_{\alpha \in \Lambda} \left| \int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \right| \\ &= \sup_{\alpha \in \Lambda} \left| \sum_{i=1}^{\ell} \frac{\int Q(z, \alpha) dF(z) - Q(z_i, \alpha)}{\ell} \right| \\ &\geq \left| \frac{\int Q(z, \alpha_\ell^*) dF(z) - Q(z_\ell, \alpha_\ell^*)}{\ell} \right| \\ &\quad - \frac{\ell - 1}{\ell} \sum_{i=1}^{\ell-1} \left| \frac{\int Q(z, \alpha_\ell^*) dF(z) - Q(z_i, \alpha_\ell^*)}{\ell - 1} \right|. \end{aligned}$$

Therefore

$$\begin{aligned}
0 &\leq \left| \frac{\int Q(z, \alpha_\ell^*) dF(z) - Q(z_\ell, \alpha_\ell^*)}{\ell} \right| \\
&\leq \sup_{\alpha \in \Lambda} \left| \int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \right| \\
&\quad + \frac{\ell-1}{\ell} \sum_{i=1}^{\ell-1} \left| \frac{\int Q(z, \alpha_\ell^*) dF(z) - Q(z_i, \alpha_\ell^*)}{\ell-1} \right|.
\end{aligned}$$

Since both terms on the right-hand side converge to zero almost surely (the first term due to the condition of the theorem and the second term due to strong law of large numbers), the nonnegative random variable on the left-hand side converges almost surely to zero.

2. Second we prove that existence of the expectation $E\xi$ of the non-negative random variable is equivalent to convergence of the sum

$$\sum_{i=1}^{\infty} P\{\xi > i\varepsilon\} < \infty \quad (3.26)$$

for any $\varepsilon > 0$.

Indeed, using the Lebesgue integral one defines expectation as the limit (for $\varepsilon \rightarrow 0$) of the sums

$$\varepsilon \sum_{i=1}^{\infty} P\{\xi > i\varepsilon\} \leq E\xi \leq \varepsilon \left(\sum_{i=1}^{\infty} P\{\xi > i\varepsilon\} + 1 \right).$$

Therefore existence of the expectation $E\xi$ is equivalent to convergence of the sum (3.26) for any $\varepsilon > 0$.

3. Third, using this fact we prove that the expectation

$$E \sup_{\alpha \in \Lambda} \left| \int Q(z, \alpha) dF(z) - Q(z_\ell, \alpha) \right| < \infty \quad (3.27)$$

exists. Indeed, note that the sequence of random variables

$$\xi_\ell = \left| \frac{\int Q(z, \alpha_\ell^*) dF(z) - Q(z_\ell, \alpha_\ell^*)}{\ell} \right|, \quad \ell = 1, 2, \dots$$

is independent. Therefore according to the Borcl–Cantelli lemma (see Chapter 1, Section 11.1), if the sequence of independent random variables ξ_ℓ converges almost surely, then the sum

$$\sum_{\ell=1}^{\infty} P \left\{ \left| \int Q(z, \alpha_\ell^*) dF(z) - Q(z_\ell, \alpha_\ell^*) \right| > \ell\varepsilon \right\} < \infty$$

is bounded. Hence Eq. (3.27) holds.

4. Furthermore, we have

$$\begin{aligned} & E \sup_{\alpha \in \Lambda} |Q(z, \alpha)| \\ & \leq E \sup_{\alpha \in \Lambda} \left| \int Q(z, \alpha) dF(z) - Q(z, \alpha) \right| + E \sup_{\alpha \in \Lambda} \left| \int Q(z, \alpha) dF(z) \right| \end{aligned} \quad (3.28)$$

Recall that we consider a set of functions satisfying constraints

$$-\infty < a \leq \int Q(z, \alpha) dF(z) \leq A < \infty.$$

Therefore from (3.27) and (3.28) we conclude that

$$E \sup_{\alpha \in \Lambda} |Q(z, \alpha)| < \infty$$

which imply the existence of an envelope.

Sublinear growth of the entropy. Now we have to prove that for any $C > 0$ the entropy of the set of functions $Q_C(z, \alpha)$, $\alpha \in A$, satisfies the required conditions. Note that if condition (3.25) is satisfied for a set of functions with some C^* , then it is satisfied for sets with $C < C^*$. Therefore it is enough to prove that condition (3.25) is satisfied for a set of functions with sufficiently large C^* .

Let us choose such large C^* that for a given small $\varepsilon > 0$ the inequality

$$\int (K(z) - C^*)_+ dF(z) < \frac{\varepsilon}{2}$$

holds true. We have

$$\begin{aligned} & \sup_{\alpha \in \Lambda} \left| \int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \right| \\ & \geq \sup_{\alpha \in \Lambda} \left| \int Q_{C^*}(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q_{C^*}(z_i, \alpha) \right| \\ & = \sup_{\alpha \in \Lambda} \int (K(z) - C^*)_+ dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} (K(z_i) - C^*)_+. \end{aligned}$$

Therefore we have

$$\begin{aligned} & \sup_{\alpha \in \Lambda} \left| \int Q_{C^*}(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q_{C^*}(z_i, \alpha) \right| \\ & \leq \sup_{\alpha \in \Lambda} \left| \int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \right| + \varepsilon + \frac{1}{\ell} \sum_{i=1}^{\ell} (K(z_i) - C^*)_+. \end{aligned}$$

The left-hand side of this expression converges almost surely to zero since the first term on the right-hand side converges almost surely to zero (according to condition of the theorem), and the last term converges almost surely to the corresponding expectation that is not larger than ε . Therefore for the uniformly bounded set of functions $Q_C(z, a), a \in A$, the uniform two-sided convergence takes place. According to Theorem 3.4, this implies condition (3.25).

3.10 KANT'S PROBLEM OF DEMARCATON AND POPPER'S THEORY OF NONFALSIFIABILITY

Thus far we have considered theorems about two-sided uniform convergence. We have constructed the characteristics of the capacity of sets of functions (which in some sense generalizes the number of functions in a finite set of functions) and then have used these characteristics (entropy of sets of indicator functions or entropy of sets of real-valued functions) to obtain the necessary and sufficient conditions for (two-sided) uniform convergence.

However, our goal is to obtain the necessary and sufficient conditions for consistency of the principle of empirical risk minimization. In Section 3.4, we showed that the condition of consistency of this induction principle coincides with the conditions of uniform one-sided convergence of means to their mathematical expectations over a given set of functions. As we shall see, the conditions for uniform one-sided convergence are expressed on the basis of conditions for uniform two-sided convergence.

However, obtaining uniform one-sided convergence using uniform two-sided convergence is not only a technical detail. To find these conditions, it is necessary to construct a mathematical generalization of one of the most impressive ideas in the philosophy of science—the idea of nonfalsifiability. In Section 3.11 we shall consider theorems about nonfalsifiability, but for now let us remind the reader what the subject of philosophy of science and the idea of nonfalsifiability are.

Since the era of ancient philosophy, two models of reasoning have been accepted:

- *deductive*, which means moving from general to particular, and
- *inductive*, which means moving from particular to general.

A model in which a system of axioms and inference rules is defined by means of which various corollaries (consequences) are obtained is ideal for the deductive approach. The deductive approach should guarantee that we obtain *true* consequences from true premises.

The inductive approach to reasoning consists of the formation of general judgments from particular assertions. However, general judgments obtained from *true* particular assertions are not always *true*. Nevertheless, it is assumed

that there exist such cases of inductive inference for which generalization assertions are justified.

The demarcation problem, originally proposed by I. Kant, is a central question of inductive theory:

What is the difference between the cases with a justified inductive step and those for which the inductive step is not justified?

The demarcation problem is usually discussed in terms of the philosophy of natural science. All theories in the natural sciences are the result of generalizations of observed real facts and therefore are built using inductive inference. In the history of natural science, there have been both true theories that reflect reality (say chemistry) and false ones (say alchemy) that do not reflect reality.

The question is the following:

Is there a formal way to distinguish between true and false theories?

Let us assume that meteorology is a true theory and astrology is a false one.

What is the formal difference between them?

- Is it in the complexity of their models?
- Is it in the predictive ability of their models?
- Is it in their use of mathematics?
- Is it in the level of formality of inference?

None of the above gives a clear advantage to either of these two theories.

The complexity of astrological models is no less than the complexity of the meteorological models.

Both theories fail in some of their predictions.

Astrologers solve differential equations for restoration of the positions of the planets, which are no simpler than the basic equations in the meteorological theory.

Finally, both theories have the same level of formalization. It contains two parts: (1) the formal description of reality and (2) the informal interpretation of it.

In the 1930s, K. Popper suggested his famous criterion for demarcation between scientific and nonscientific theories.[†] According to Popper, a necessary condition for justifiability of a theory is the feasibility of its falsification. By the falsification of a theory, Popper means the existence of a collection of

[†]Popper used the terminology of empirical and metaphysical theories.

particular assertions which cannot be explained by the given theory although they fall into its domain. If the given theory can be falsified, it satisfies the necessary conditions of a scientific theory.

Let us come back to our example. Both meteorology and astrology make weather forecasts. Consider the following assertion:

in the New York area, both a tropical storm and snowfall can happen in one hour.

Suppose that according to the theory of meteorology, this is impossible. Then this assertion falsifies the theory because if such a situation really will happen (note that nobody can guarantee with probability one that this is impossible) the theory will not be able to explain it. In this case the theory of meteorology satisfies the necessary conditions to be viewed as a scientific theory.

Suppose that this assertion can be explained by the theory of astrology. (There are many elements in the starry sky, and they can be used to create an explanation.) In this case, this assertion does not falsify the theory. If there is no example that can falsify the theory of astrology, then astrology according to Popper should be considered a nonscientific theory.

In the next section we describe the theorems of nonfalsifiability. We show that if for some set of functions, conditions for uniform convergence do not hold, the situation of nonfalsifiability will arise.

3.1.1 THEOREMS ABOUT NONFALSIFIABILITY

In this section we show that if uniform two-sided convergence does not take place, then the method of empirical risk minimization is nonfalsifiable.

3.1.1.1 Case of Complete Nonfalsifiability

To give a clear explanation of why this happens, let us start with the simplest case. Suppose for the set of indicator functions $Q(z, \alpha), \alpha \in A$, the following equality is true:

$$\lim_{\ell \rightarrow \infty} \frac{H^A(\ell)}{\ell} = \ln 2. \quad (3.29)$$

Intuitively, it is clear that the ratio of the entropy to the number of observations $H^A(\ell)/\ell$ monotonically decreases when the number of observations ℓ increases. (This is proven formally in Chapter 14.) Thus, if (3.29) happened, then for any finite number ℓ the equality

$$\frac{H^A(\ell)}{\ell} = \ln 2$$

holds true.

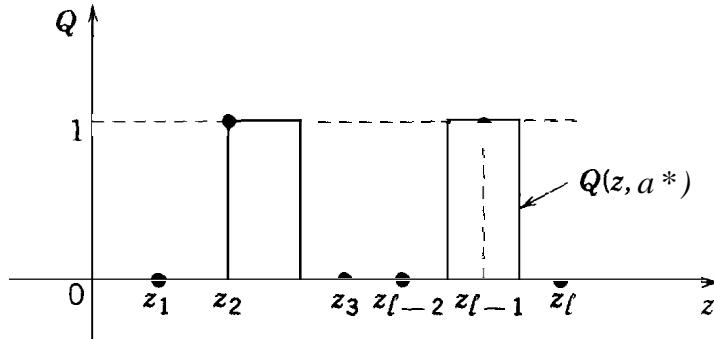


FIGURE 3.6. A learning machine with the set of functions $Q(z, \alpha)$, $\alpha \in A$, is **nonfalsifiable** if for almost all samples z_1, \dots, z_ℓ given by the generator of examples and for any possible labels $\delta_1, \dots, \delta_\ell$ for these z 's, the machine contains a function $Q(z, a^*)$ that provides equalities $\delta_i = Q(z_i, a)$, $i = 1, \dots, \ell$.

According to the definition of entropy, this means that for almost all samples z_1, \dots, z_ℓ the equality

$$N^\Lambda(z_1, \dots, z_\ell) = 2^\ell$$

is valid.

In other words, the set of functions of the learning machine is such that almost any sample z_1, \dots, z_ℓ (of arbitrary size ℓ) can be separated in all possible ways by functions of this set. This implies that the minimum of empirical risk for this machine equals zero. We call this learning machine nonfalsifiable because it can give a general explanation (function) for almost any data (see Fig. 3.6).

3.11.2 Theorem About Partial Nonfalsifiability

In the case when entropy of the set of indicator functions over the number of observations tends to a nonzero limit, the following theorem shows that there exists some subspace of the original space Z where the learning machine is nonfalsifiable.

Theorem 3.6. *For the set of indicator functions $Q(z, a)$, $a \in A$, let the convergence*

$$\lim_{\ell \rightarrow \infty} \frac{H^\Lambda(\ell)}{\ell} = c > 0$$

be valid.

Then there exists a subset Z^ of the set Z such that*

$$(a) \quad P(Z^*) = c$$

and (b) for the subset

$$z_1^*, \dots, z_k^* = (z_1, \dots, z_\ell) \cap Z^*$$

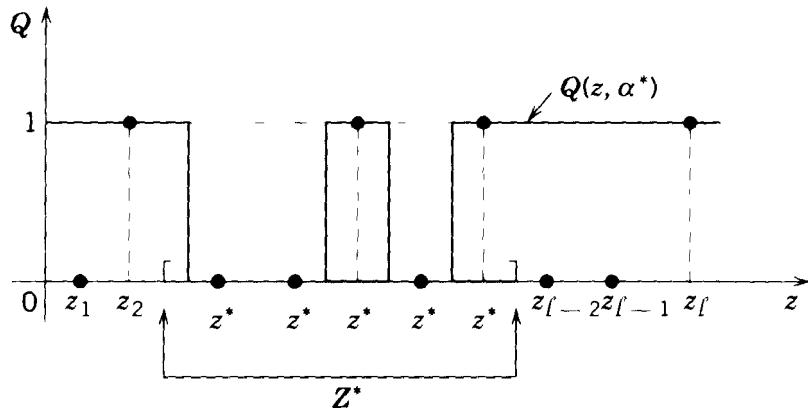


FIGURE 3.7. A learning machine with the set of functions $Q(z, \alpha), \alpha \in A$, is partially *nonfalsifiable* if there exists a region $Z^* \in Z$ with nonzero measure such that for almost all samples z_1, \dots, z_l , given by the generator of examples and for any labels $\delta_1, \dots, \delta_l$, for these z 's, the machine contains a function $Q(z, \alpha')$ that provides equalities $\delta_i = Q(z_i, \alpha)$ for all z_i belonging to the region Z^* .

of almost any training set

$$z_1, \dots, z_l$$

that belongs to Z^* and for any given sequence of binary values

$$\delta_1, \dots, \delta_k, \quad \delta_i \in \{0, 1\}$$

there exists a function $Q(z, \alpha^*)$ for which the equalities

$$\delta_i = Q(z_i^*, \alpha^*), \quad i = 1, 2, \dots, k$$

hold true.

This theorem shows that if conditions of uniform convergence fail, then there exists some subspace of the input space where the learning machine is nonfalsifiable (see Fig. 3.7).

3.11.3 Theorem About Potential Nonfalsifiability

Now let us consider the set of uniformly bounded real-valued functions

$$|Q(z, \alpha)| \leq C, \quad \alpha \in \Lambda.$$

For this set of functions a more sophisticated model of nonfalsifiability will be used. So we give the following definition of nonfalsifiability:

Definition. We say that the learning machine that has an admissible set of real-valued functions $Q(z, \alpha), \alpha \in A$, is *potentially nonfalsifiable* if there exist

two functions^t

$$\psi_1(z) \geq \psi_0(z)$$

such that:

1. There exists some positive constant c for which the equality

$$\int (\psi_1(z) - \psi_0(z)) dF(z) = c > 0$$

holds true.

2. For almost any sample

$$z_1, \dots, z_\ell,$$

any sequence of binary values,

$$\delta_1, \dots, \delta_\ell, \quad \delta_i \in \{0, 1\},$$

and any ε , one can find a function $Q(z, \alpha^*)$ in the set of functions $Q(z, \alpha), \alpha \in A$, for which the inequalities

$$|\psi_{\delta_i}(z_i) - Q(z_i, \alpha^*)| < \varepsilon, \quad \delta_i \in \{0, 1\}$$

hold true.

In this definition of nonfalsifiability, we use two essentially different functions $\psi_1(z)$ and $\psi_0(z)$ to generate the values y_i of the function for the given vectors z_i . To make these values arbitrary, one can switch between these two functions using the arbitrary rule δ_i . The set of functions $Q(z, \alpha), \alpha \in A$, forms a potentially nonfalsifiable machine if for almost any sequence of pairs $(\psi_{\delta(i)}(z_i), z_i)$ obtained on the basis of random vectors z_i and this switching rule $\delta(i)$, one can find in this set a function $Q(z, \alpha^*)$ that describes these pairs with high accuracy (Fig. 3.8).

Note that this definition of nonfalsifiability generalizes Popper's notion. In our simplest example considered in the beginning of Section 3.10, for the set of indicator functions $Q(z, \alpha), \alpha \in A$, we use this notion of nonfalsifiability where $\psi_1(z) = 1$ and $\psi_0(z) = 0$. In Theorem 3.6 we actually use the functions

$$\psi_1(z) = \begin{cases} 1 & \text{if } z \in Z^*, \\ Q(z, \bar{\alpha}) & \text{if } z \notin Z^*, \end{cases} \quad \psi_0(z) = \begin{cases} 0 & \text{if } z \in Z^*, \\ Q(z, \bar{\alpha}) & \text{if } z \notin Z^*, \end{cases}$$

where $Q(z, \bar{\alpha})$ is some function from the given set of indicator functions.

On the basis of this concept of potential nonfalsifiability, we formulate the following theorem that holds for an arbitrary set of uniformly bounded functions (including sets of indicator functions).

^tThese two functions need not necessarily belong to the set $Q(z, \alpha), \alpha \in A$.

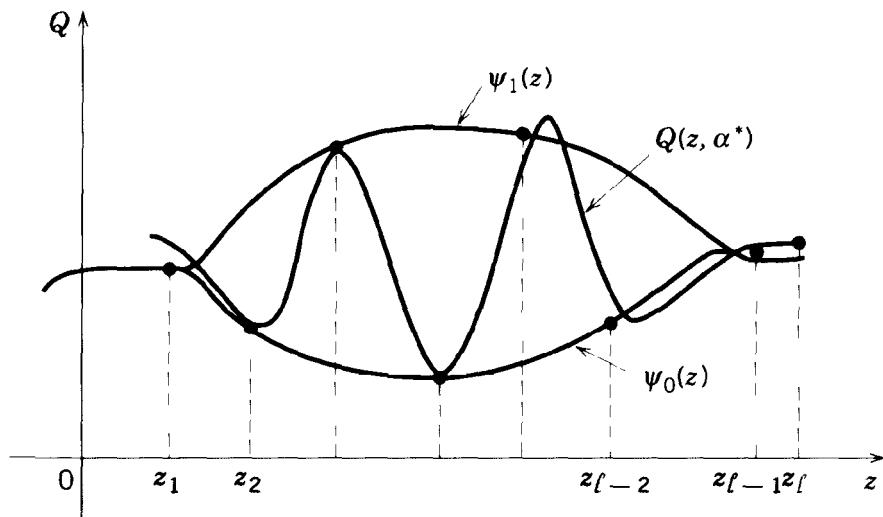


FIGURE 3.8. A learning machine with the set of functions $Q(z, a)$, $a \in A$, is potentially nonfalsifiable if for any $\varepsilon > 0$ there exist two essentially different functions $\psi_1(z)$ and $\psi_0(z)$ such that for almost all samples z_1, \dots, z_ℓ given by the generator of examples, and for any values u_1, \dots, u_ℓ constructed on the basis of these curves using the rule $u_i = \psi_{\delta_i}(z_i)$, where $\delta_i \in \{0, 1\}$ is an arbitrary binary function, the machine contains a function $Q(z, \alpha^*)$ that satisfy inequalities $|\psi_{\delta_i}(z_i) - Q(z_i, \alpha^*)| \leq \varepsilon$, $i = 1, \dots, \ell$.

Theorem 3.7. Suppose that for the set of uniformly bounded real-valued functions $Q(z, a)$, $a \in A$, there exist ϵ_0 such that the convergence

$$\lim_{\ell \rightarrow \infty} \frac{H^\Lambda(\epsilon_0, \ell)}{\ell} = c^* > 0$$

is valid.

Then the learning machine with this set of functions is potentially nonfalsifiable.

This theorem will be proved in Chapter 16.

3.12 CONDITIONS FOR ONE-SIDED UNIFORM CONVERGENCE AND CONSISTENCY OF THE EMPIRICAL RISK MINIMIZATION PRINCIPLE

We start this section with an example of a learning machine that has a set of functions which make it nonfalsifiable, but, nevertheless, the machine can generalize using the empirical risk minimization principle. This happens because learning theory considers the nonsymmetric situation where the machine must generalize by minimizing risk rather than by maximizing risk.

Example. Let $z \in (0,1)$, and let $F(z)$ be a uniform distribution function. Consider the following set of two parametric indicator functions $Q(z, \alpha, \beta)$,

$\mathbf{a} \in \mathbf{A} = (0, 1)$, $\beta \in \mathcal{B}$: $Q(z, \mathbf{a}, \beta) = 1$, for $z \geq \mathbf{a}$ and $Q(z, \mathbf{a}, \beta) = 0$ for all $z < \mathbf{a}$ except for a finite number of points where it equals 1. This specific finite number of points is determined by the parameter β . The set of functions is such that for any finite set of points in the region $(0, 1)$ there exists a function specified by the parameter $\beta \in \mathcal{B}$ which takes the value of one at these points. It is easy to see that the learning machine that contains this set of functions is nonfalsifiable (see Fig. 3.9). Indeed, for any set of (different) vectors

$$z_1, \dots, z_\ell$$

and any sequence

$$\delta_1, \dots, \delta_\ell, \quad \delta_i \in \{0, 1\}$$

there exist parameters $a = a^*$ and $\beta = \beta^*$ which provide equalities

$$\delta_i = Q(z_i, \alpha^*, \beta^*), \quad i = 1, 2, \dots, \ell$$

For this set of functions, the equality

$$\frac{H^{\Lambda, \mathcal{B}}(\ell)}{\ell} = \frac{\int \ln N^{\Lambda, \mathcal{B}}(z_1, \dots, z_\ell) dz_1, \dots, dz_\ell}{\ell} = \ln 2$$

is valid.

Note that the value of the risk functional

$$R(\alpha, \beta) = \int Q(z, \alpha, \beta) dz$$

depends on \mathbf{a} and does not depend on β .

Consider another learning machine that contains the following set of functions:

$$Q^*(z, \bar{\alpha}) = \begin{cases} 0 & \text{if } z < \bar{\alpha} \\ 1 & \text{if } z \geq \bar{\alpha} \end{cases} \quad \bar{\alpha} \in [0, 1].$$

Now, suppose that both of our learning machines use the empirical risk minimization induction principle and the same training set

$$z_1, \dots, z_\ell.$$

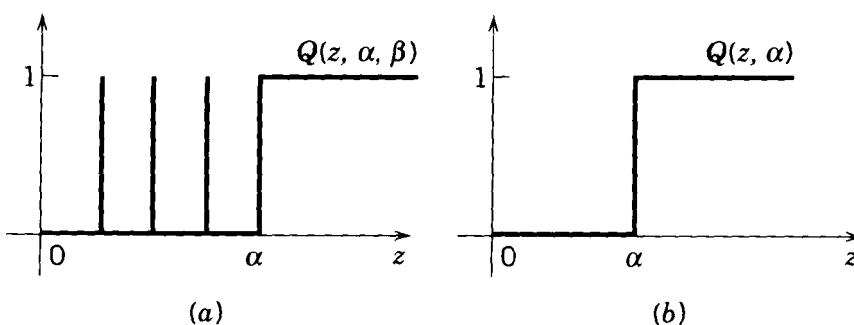


FIGURE 3.9. Two learning machines, one nonfalsifiable (with set of functions in part a) and another falsifiable (with set of functions in part b), provide the same results.

It is clear that for any function $Q(z, \alpha, \beta)$ of the first machine, there exists a function $Q^*(z, \bar{\alpha})$ of the second machine such that

$$\frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha, \beta) \geq \frac{1}{\ell} \sum_{i=1}^{\ell} Q^*(z_i, \bar{\alpha}),$$

$$\int Q(z, \alpha, \beta) dz = \int Q^*(z, \bar{\alpha}) dz.$$

But, according to the Glivenko–Cantelli theorem for the class of functions used in the second machine, uniform convergence takes place (see Chapter 2, Section 2.4). If for the second machine uniform convergence takes place, then for the first machine one-sided uniform convergence takes place. According to Theorem 3.1, this implies the consistency of the learning machine using the empirical risk minimization principle.

This example is important because it describes the general idea when such a situation is possible. Let us repeat once more the idea of this example. We considered the set of real-valued functions $Q(z, \alpha, \beta)$, $\alpha \in A, \beta \in B$, for which (two-sided) uniform convergence does not take place. Then we introduced a new set of functions $Q^*(z, \bar{\alpha})$, $\bar{\alpha} \in A$, which had the following property: For any function $Q(z, \alpha, \beta)$ in the first set there was a function $Q^*(z, \bar{\alpha})$ in the second set such that

$$Q(z, \alpha, \beta) \geq Q^*(z, \bar{\alpha})$$

$$\int (Q(z, \alpha, \beta) - Q^*(z, \bar{\alpha})) dF(z) < \varepsilon \quad (3.30)$$

(in the example $F(z) = z$), where ε is a arbitrary small value. We used the fact that if for the second set of functions uniform convergence was valid, then for the first set of functions one-sided uniform convergence takes place (Fig. 3.10).

Exactly this scheme of reasoning will be repeated in the theorem about one-sided uniform convergence. Let us consider a set of uniformly bounded functions $Q(z, \alpha)$, $\alpha \in A$. We assume that all constructions we used to prove the theorem are measurable with respect to the distribution function $F(z)$.

Theorem 3.8. *For uniform one-sided convergence to take place on a set of uniformly bounded functions $Q(z, \alpha)$, $\alpha \in A$, it is necessary and sufficient that for any positive ϵ, δ , and ε , there exists a set of functions $Q^*(z, \bar{\alpha})$, $\bar{\alpha} \in A$, such that (see Fig. 3.10):*

1. *For any function $Q(z, \alpha)$ there exists a function $Q^*(z, \bar{\alpha})$ satisfying the conditions*

$$Q(z, \alpha) \geq Q^*(z, \bar{\alpha})$$

$$\int (Q(z, \alpha) - Q^*(z, \bar{\alpha})) dF(z) < \varepsilon.$$

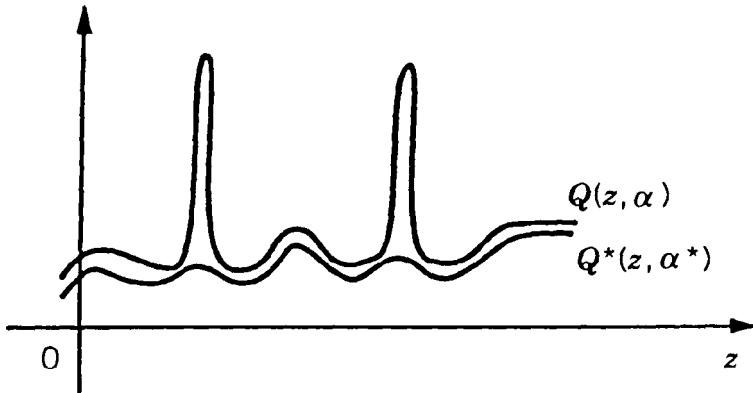


FIGURE 3.10. For any function $Q(z, \alpha)$, $\alpha \in A$, one considers a function $Q^*(z, \alpha^*)$, $\alpha^* \in A^*$, such that $Q^*(z, \alpha^*)$ does not exceed $Q(z, \alpha)$ and is close to it.

2. The ϵ -entropy of the set of functions $Q^*(z, \bar{\alpha})$, $\bar{\alpha} \in A$, satisfies the inequality

$$\lim_{\ell \rightarrow \infty} \frac{H^\Delta(\epsilon, \ell)}{\ell} < \delta. \quad (3.31)$$

Remark. This theorem gives necessary and sufficient conditions for one-sided uniform convergence for some fixed probability measure $F(z)$. In order that uniform convergence take place for any probability measure $F \bullet P$ it is necessary and sufficient that inequality (3.31) be valid for any $F \bullet P$.

Chapter 16 is devoted to the proving this theorem. As we shall see, to prove the sufficient conditions of this theorem, we use the same technique as we use for proving sufficient conditions for two-sided uniform convergence in Theorem 3.4. This technique is actually based on the same idea which in three lines gives the result for the Simplest Model (Section 3.6). The essential difference, however, is that instead of a number of functions in set N , we use the entropy $H^\Delta(\epsilon, \ell)$.

The main difficulties in proving this theorem arise in proving the necessity of the conditions (3.31). The proof of the necessity of these conditions is based on the theorem about potential nonfalsifiability and will be done in three steps:

1. First, we shall derive the following necessary (but not sufficient) conditions:

Theorem 3.9. For one-sided uniform convergence to take place, it is necessary that for any ϵ there should exist a finite ϵ -net of the set $Q(z, \alpha)$, $\alpha \in A$ in the metric $L_1(P)$:

$$\rho(\alpha_1, \alpha_2) = \int |Q(z, \alpha_1) - Q(z, \alpha_2)| dF(z).$$

2. Next we shall prove that if the learning machine with a set of functions $Q^*(z, a)$, $a \in A$, satisfying (3.30) is potentially nonfalsifiable, then there exist in the set $Q(z, a)$, $a \in A$, two functions $Q(z, a^*)$ and $Q(z, a_*)$ which are ε -close to the functions $\psi_1(z)$ and $\psi_0(z)$ in the metric $L_1(P)$. For these functions, the inequality

$$\int |Q(z, a^*) - Q(z, a_*)| dF(z) > c - 2\varepsilon, \quad c > 0$$

holds true.

3. Using these two facts, we prove the necessity of the conditions (3.31) by the following reasoning.

We assume that one-sided uniform convergence takes place; and at the same time for the set of functions satisfying (3.30), condition (3.31) fails. This will bring us to a contradiction. On the one hand since uniform one-sided convergence holds, there exists a finite ε -net (Theorem 3.9) and therefore the distance in $L_1(F)$ between any functions within one element of the ε -net is less than 2ε .

On the other hand since condition (3.31) does not hold, there exists among the elements of a finite ε -net at least one that contains functions for which condition (3.31) does not hold. Since the machine that contains functions of this element is potentially nonfalsifiable, it has two functions with distance larger than $c - 2\varepsilon$. Appropriate choices of c and ε give the contradiction.

Thus, Theorem 3.8 gives the necessary and sufficient conditions for uniform one-sided convergence. According to the corollary to Theorem 3.1, these conditions are equivalent to the necessary and sufficient conditions for consistency of the learning machine which uses the empirical risk minimization induction principle.[†] This theorem, therefore, completes the theory of consistency of the learning processes.

However, to complete the conceptual model of the learning theory we have to answer two additional questions.

Theorem 3.8 determines the conditions when the learning machine is consistent. However, it says nothing about the rate of convergence of the obtained risks $R(\alpha_t)$ to the minimal risk $R(\alpha_0)$. It is possible to construct examples where the ERM principle is consistent but has an arbitrary slow rate of convergence.

The fundamental questions are:

1. *What are the conditions for the existence of a fast (with exponential*

[†]This theorem also gives the necessary and sufficient conditions for the consistency of the maximum likelihood method in the case when the set of densities is uniformly bounded and uniformly separated from zero (see Theorem 3.2 in Section 3.6).

bounds) asymptotic rate of uniform convergence for a given probability measure?

To answer this question means to describe the conditions under which there exist two positive constants b and c such that for sufficiently large $\ell > \ell(\varepsilon, A, P)$, the inequality

$$P \left\{ \sup_{\alpha \in \Lambda} \left| \int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \right| > \varepsilon \right\} < b \exp\{-c\varepsilon^2\ell\} \quad (3.32)$$

holds true.

2. *What are the conditions for existence of a fast asymptotic rate of uniform convergence for any probability measure $F(z) \bullet \mathcal{P}_0$?*

To answer this question means to describe the necessary and sufficient conditions under which there exist two positive constants b and c such that for sufficiently large $\ell > \ell(\varepsilon, A)$ the inequality

$$\sup_{F(z) \in \mathcal{P}_0} P \left\{ \sup_{\alpha \in \Lambda} \left| \int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \right| > \varepsilon \right\} < b \exp\{-c\varepsilon^2\ell\} \quad (3.33)$$

holds true. Note that this question for the set of indicator functions $Q(z, a), a \in A$, forms the general Glivenko-Cantelli problem[†] (see Chapter 2, Section 2.4).

In the subsequent chapters we shall give in detail the answers to both questions.

These answers will be based on some fundamental concepts of capacity of a set of functions implemented by the learning machine. These concepts are constructed on the basis of the concept entropy of a set of functions for the sample of size ℓ , considered in this chapter.

3.13 THREE MILESTONES IN LEARNING THEORY

The most important result of the described conceptual part of the learning theory is the fact that the introduced capacity concept (the entropy) completely defines the *qualitative* behavior of the learning processes: the consistency of learning. As we will see, the robust characteristics of this concept

[†]The Generalized Glivenko-Cantelli problem introduced in Chapter 2 considers convergence in probability uniformly for all probability measures. However, if this convergence takes place, the bound (3.32) is valid as well.

define *quantitative* singularity of learning processes as well: the nonasymptotic bound on the rate of convergence of the learning processes for both the distribution-dependent and the distribution-independent cases,

Obtaining these bounds is the subject of Chapter 4 and Chapter 5. The goal of this last section is to define the structure of capacity concepts that we use in this book and demonstrate their connections.

For simplicity, we first consider the set of indicator functions $Q(z, a), a \in A$ (i.e., the problem of pattern recognition), and then consider the set of real-valued functions.

As mentioned above, in the case of indicator functions $Q(z, a), a \in A$, the minimal ε -net of the vectors $q(\alpha), \alpha \in A$ (see Section 2.3.3), does not depend on ε if $\varepsilon < 1$. The number of elements in the minimal ε -net

$$N^A(z_1, \dots, z_\ell) = N^\Lambda(\varepsilon; z_1, \dots, z_\ell)$$

is equal to the number of different separations of the data z_1, \dots, z_ℓ by functions of the set $Q(z, a), a \in A$.

For this set of functions the entropy also does not depend on ε :

$$H^\Lambda(\ell) = E \ln N^A(z_1, \dots, z_\ell),$$

where expectation is taken over (z_1, \dots, z_ℓ) .

Consider two new concepts that are constructed on the basis of the values of $N^A(z_1, \dots, z_\ell)$:

1. The *annealed entropy*

$$H_{\text{ann}}^\Lambda(\ell) = \ln E N^A(z_1, \dots, z_\ell);$$

2. The *growth function*

$$G^\Lambda(\ell) = \ln \sup_{z_1, \dots, z_\ell} N^A(z_1, \dots, z_\ell).$$

These concepts are defined in such a way that for any ℓ the inequalities

$$H^\Lambda(\ell) \leq H_{\text{ann}}^\Lambda(\ell) \leq G^\Lambda(\ell)$$

are valid.

Indeed, the first inequality immediately comes from applying the Jensen inequality to the entropy (for Jensen's inequality see Chapter 1, Eq. 1.12), the second inequality is obvious. On the basis of these functions the main milestones of learning theory are constructed.

In Theorem 3.3 we introduce the condition

$$\lim_{\ell \rightarrow \infty} \frac{H^\Lambda(\ell)}{\ell} = 0 \quad (3.34)$$

describing a *sufficient condition* for consistency of the ERM principle (the necessary and sufficient conditions are given by a slightly different condition described in Theorem 3.8). This equation is the *first milestone* in the pattern recognition theory: We require that any machine minimizing empirical risk should satisfy it.

However, this equation says nothing about the rate of convergence of the obtained risks $R(\alpha_\ell)$ to the minimal one $R(\alpha_0)$. It is possible to construct examples where the ERM principle is consistent, but where the risks have arbitrarily slow asymptotic rate of convergence.

It turns out that the equation

$$\lim_{\ell \rightarrow \infty} \frac{H_{\text{ann}}^\Lambda(\ell)}{\ell} = 0 \quad (3.35)$$

is a *sufficient* condition for a fast rate of convergence[†] defined by condition (3.32). This equation is the *second milestone* in the pattern recognition theory: It guarantees a fast asymptotic rate of convergence.

Thus far, we have considered two equations: One equation describes the necessary and sufficient condition for the consistency of the ERM method, and the other describes the sufficient condition for fast rate of convergence of the ERM method. Both equations are valid for a *given* probability measure $F(z)$ on the observations (both the entropy $H^\Lambda(\ell)$ and the annealed entropy $H_{\text{ann}}^\Lambda(\ell)$ are constructed using this measure). However, our goal is to construct a learning machine capable of solving many different problems (for many different probability measures).

The following equation describes the *necessary and sufficient conditions* for consistency of ERM for *any* probability measure:

$$\lim_{\ell \rightarrow \infty} \frac{H^\Lambda(\ell)}{\ell} = 0. \quad (3.36)$$

It is also the case that if this condition holds true, then the rate of convergence is fast.

This equation is the *third milestone* in the pattern recognition theory. It describes the necessary and sufficient condition under which a learning machine that implements the ERM principle has high asymptotic rate of convergence independent of the probability measure (i.e., independent of the problem that has to be solved).

In more general case when we consider bounded real-valued functions the necessary and sufficient conditions for consistency of empirical risk minimization method is dependent on ε entropy

$$H^\Lambda(\varepsilon; \ell) = E \ln N^\Lambda(\varepsilon; z_1, \dots, z_\ell)$$

[†]The necessity of this condition for a fast rate of convergence is an open question.

(for simplicity we consider the minimal ε -net in C-metric). According to Theorem 3.4 the convergence

$$\frac{H^\Lambda(\varepsilon; \ell)}{\ell} \xrightarrow[\ell \rightarrow \infty]{} 0, \quad \forall \varepsilon > 0$$

defines the sufficient condition for consistency of learning processes (the slightly different condition given in Theorem 3.8 defines the necessary and sufficient conditions).

This equality is the first milestone in the learning theory.

In Chapter 15 we prove Theorem 15.2, which states that the fast rate of convergence of a learning process is valid if the *annealed ε -entropy*

$$H_{\text{ann}}^\Lambda(\varepsilon; \ell) = \ln E N^\Lambda(\varepsilon; z_1, \dots, z_\ell)$$

is such that convergence

$$\frac{H_{\text{ann}}^\Lambda(\varepsilon; \ell)}{\ell} \xrightarrow[\ell \rightarrow \infty]{} 0, \quad \forall \varepsilon > 0$$

takes place.

This equality is the second milestone in learning theory.

Lastly, consider the growth function

$$G^\Lambda(\varepsilon; \ell) = \ln \sup_{z_1, \dots, z_\ell} N^\Lambda(\varepsilon; z_1, \dots, z_\ell).$$

The equation

$$\frac{G^\Lambda(\varepsilon; \ell)}{\ell} \xrightarrow[\ell \rightarrow \infty]{} 0, \quad \forall \varepsilon > 0$$

describes the condition under which the learning process is consistent and has a fast rate of convergence for any probability measure. This equation is the third milestone in the learning theory.

These milestones form the cornerstones for constructing bounds for the rate of convergence of learning machines which we consider in Chapters 4 and 5.

4

BOUNDS ON THE RISK FOR INDICATOR LOSS FUNCTIONS

Beginning with this chapter we start to study the rate of convergence of the learning processes. We look for the bounds that estimate two quantities:

1. The value of achieved risk for the function minimizing the empirical risk.
2. The difference between the value of achieved risk and the value of minimal possible risk for a given set of functions.

These bounds determine generalization ability of the learning machines utilizing the empirical risk minimization induction principle.

In this chapter we consider the special set of loss functions, namely, the set of indicator functions (that are specific for the pattern recognition problem). Our goal is to obtain the bounds on the rate of uniform convergence of frequencies to their probabilities over a given set of events (defined by indicator functions).

Deriving two types of bounds constitutes the main contents of this chapter:

1. Bounds on the rate of uniform convergence
2. Bounds on the rate of relative uniform convergence

To obtain these bounds we use two capacity concepts introduced in Chapter 3: the annealed entropy and the growth function. Using these concepts we derive both distribution-dependent bounds and distribution-independent bounds. These bounds, however, are nonconstructive since the theory does not provide us with clear methods to evaluate them in specific cases. Therefore we consider a new capacity concept: the VC dimension of a set of functions that can be evaluated for any given set of functions.

In terms of the VC dimension we obtain constructive distribution-independent bounds on the rate of uniform convergence.

4.1 BOUNDS FOR THE SIMPLEST MODEL: PESSIMISTIC CASE

Consider the problem of minimizing the risk functional

$$R(\alpha) = \int Q(z, \alpha) dF(z), \quad \alpha \in \Lambda \quad (4.1)$$

on the basis of empirical data

$$z_1, \dots, z_\ell, \quad (4.2)$$

where $Q(z, \alpha)$, $\alpha \in \Lambda$, is a set of indicator functions.

To minimize risk (4.1) on the basis of data (4.2) we use the principle of empirical risk minimization. Instead of (4.1) we minimize the empirical risk functional

$$R_{\text{emp}}(\alpha) = \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha), \quad \alpha \in \Lambda \quad (4.3)$$

over the set of functions $Q(z, \alpha)$, $\alpha \in \Lambda$. For the indicator functions the risk (4.1) describes the probability of events $A_i = \{z : Q(z, \alpha) = 1\}$, $\alpha \in \Lambda$, and the empirical risk functional (4.3) describes the frequency of these events.

Suppose the minimum of the risk functional (4.1) is achieved on the function $Q(z, \alpha_0)$ and the minimum of the empirical risk functional (4.3) is achieved on the function $Q(z, \alpha_\ell)$, $\alpha_\ell = \alpha(z_1, \dots, z_\ell)$.

To estimate the generalization ability of the principle of empirical risk minimization we have to answer two questions:

- ***What value of the risk does the function $Q(z, \alpha_\ell)$ provide?***
To answer this question means to estimate the value $R(\alpha_\ell)$.
- ***How close is this risk to the smallest possible for a given set of firnctions?***
To answer this question means to estimate the difference

$$\Delta(\alpha_\ell) = R(\alpha_\ell) - R(\alpha_0).$$

The answers to both these questions are based on the study of the rate of uniform convergence

$$\sup_{\alpha \in \Lambda} \left(\int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \right) \xrightarrow[\ell \rightarrow \infty]{P} 0.$$

We start our studies with the simplest model which we have already met in Chapter 3 (Section 3.7.2).

4.1.1 The Simplest Model

We consider the model where a *set of indicator functions contains a finite number N of elements* $Q(z, \alpha_k)$, $k = 1, 2, \dots, N$.

In this section and in the next two sections we shall estimate the rate of uniform convergence for the simplest model. We shall obtain the rate of convergence which depends on the capacity of a set of functions (logarithm of the number N of functions in a set).

The main goal of this chapter is the generalization of the results obtained for sets with a finite number of functions to sets of functions that contain an infinite number of elements. To get this generalization we shall introduce appropriate concepts of capacity of the set of functions, and then in terms of these concepts we will obtain expressions for the bounds of the rate of uniform convergence. These expressions are similar to ones derived for the Simplest Model.

Below we will use *additive Chernoff bounds* which are valid for the 4 random independent trials in the Bernoulli scheme:

$$P \{ p - \nu_\ell > \varepsilon \} < \exp \left\{ -2\varepsilon^2 \ell \right\}, \quad (4.4)$$

$$P \{ \nu_\ell - p > \varepsilon \} < \exp \left\{ -2\varepsilon^2 \ell \right\}. \quad (4.5)$$

To estimate the rate of uniform convergence we consider the sequence of inequalities

$$\begin{aligned} & P \left\{ \sup_{1 \leq k \leq N} \left(\int Q(z, \alpha_k) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha_k) \right) > \varepsilon \right\} \\ & \leq \sum_{k=1}^N P \left\{ \left(\int Q(z, \alpha_k) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha_k) \right) > \varepsilon \right\} \\ & \leq N \exp \{-2\varepsilon^2 \ell\}. \end{aligned} \quad (4.6)$$

To get (4.6) we use Chernoff inequality (4.4) (recall that for indicator functions the risk functional defines probabilities and the empirical risk functional defines frequency).

Let us rewrite this inequality in the equivalent form. To do this we introduce a positive value $0 < \eta \leq 1$ and the equality

$$N \exp \{-2\varepsilon^2 \ell\} = \eta$$

which we solve with respect to ε . We obtain

$$\varepsilon = \sqrt{\frac{\ln N - \ln \eta}{2\ell}}. \quad (4.7)$$

Now the assertion (4.6) has the following equivalent form:

With probability $1 - \eta$ simultaneously for all N functions in the set $Q(z, \alpha_k)$, $k = 1, 2, \dots, N$, the inequality

$$\int Q(z, \alpha_k) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha_k) \leq \sqrt{\frac{\ln N - \ln \eta}{2\ell}} \quad (4.8)$$

is valid.

Let $Q(z, \alpha_{k(0)})$ be a function from our finite set of functions that minimizes the risk (4.1) and let $Q(z, \alpha_{k(\ell)})$ be a function from this set that minimizes the empirical risk (4.3). Since the inequality (4.8) is true for all functions in the set, it is true as well for the function $Q(z, \alpha_{k(\ell)})$.

Thus with probability at least $1 - \eta$ the following inequality

$$\int Q(z, \alpha_{k(\ell)}) dF(z) \leq \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha_{k(\ell)}) + \sqrt{\frac{\ln N - \ln \eta}{2\ell}} \quad (4.9)$$

is valid.

This inequality estimates the value of the risk for the chosen function $Q(z, \alpha_{k(\ell)})$. It answers the first question about the generalization ability of the principle of empirical risk minimization for the simplest model.

To answer the second question (how close is the risk for the chosen function to the minimal one), note that for the function $Q(z, \alpha_{k(0)})$ which minimizes the expected risk (4.1) the Chernoff inequality

$$P \left\{ \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha_{k(0)}) - \int Q(z, \alpha_{k(0)}) dF(z) > \varepsilon \right\} \leq \exp\{-2\varepsilon^2\ell\} \quad (4.10)$$

holds true.

This inequality implies that with probability $1 - \eta$ the inequality

$$\int Q(z, \alpha_{k(0)}) dF(z) \geq \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha_{k(0)}) - \sqrt{\frac{-\ln \eta}{2\ell}} \quad (4.11)$$

holds true.

Since $Q(z, \alpha_\ell)$ minimizes empirical risk the inequality

$$\sum_{i=1}^{\ell} Q(z_i, \alpha_{k(0)}) - \sum_{i=1}^{\ell} Q(z_i, \alpha_{k(\ell)}) \geq 0$$

is valid. Taking this inequality into account from (4.9) and (4.11) we conclude that with probability at least $1 - 277$ the inequality

$$\Delta(\alpha_{k(\ell)}) = R(\alpha_{k(\ell)}) - R(\alpha_{k(0)}) \leq \sqrt{\frac{\ln N - \ln \eta}{2\ell}} + \sqrt{\frac{-\ln \eta}{2\ell}} \quad (4.12)$$

holds true.

Thus the two inequalities, namely (4.9) and (4.12), give complete information about the generalization ability of the method of empirical risk minimization for the case when a set of functions contains a finite number of elements: Inequality (4.9) estimates the upper bound of the risk for chosen function, and inequality (4.12) estimates how close is this bound of the risk to the minimal possible risk for this set of functions.

Note that these bounds are tight. In general it is impossible to improve the right-hand side of inequalities (4.9) or (4.12). Indeed, consider the case when $N = 1$. Let the function $Q(z, \alpha_1)$ be such that the risk $R(\alpha_1)$ (probability of the event $\{z : Q(z, \alpha_1) = 1\}$; let us call this event the error) is close to $1/2$. In this case the empirical risk describes the frequencies of error ν_ℓ estimated in the Bernoulli scheme with ℓ trials. When ℓ is rather large the following approximation

$$P\{p - \nu_\ell > \varepsilon\} \sim \exp\{-2\varepsilon^2\ell\}$$

is quite tight.

Thus the inequalities (4.9), (4.12) cannot be improved if a *set of functions contains only bad functions (that provide probability of error close to 1/2)*.

4.2 BOUNDS FOR THE SIMPLEST MODEL: OPTIMISTIC CASE

However, the situation changes dramatically if a set of functions contains at least one good function (which provides probability of error equal to zero). Suppose that among our N functions there exists at least one with zero probability of error. Then in accordance with the principle of empirical risk minimization, one should choose the function which provides zero error on a given sample. It is possible that there exist several such functions. Let us choose any of them.

What is the probability that the function that provides zero empirical risk has the expected risk larger than a given positive constant ε ? To estimate this probability, one has to bound the expression

$$P \left\{ \sup_{1 \leq k \leq N} |R(\alpha_k) - R_{\text{emp}}(\alpha_k)| \hat{\theta}(R_{\text{emp}}(\alpha_k)) > \varepsilon \right\},$$

where $R(\alpha_k)$ is the value of the expected risk (4.1) for the function $Q(z, \alpha_k)$ and $R_{\text{emp}}(\alpha_k)$ is the value of the empirical risk (4.3) for this function, and $\hat{\theta}(R_{\text{emp}}(\alpha_k))$ is the following indicator:

$$\hat{\theta}(u) = \begin{cases} 1 & \text{if } u = 0, \\ 0 & \text{if } u > 0. \end{cases}$$

Let us bound this probability. The following sequence of inequalities is valid for $N > 1$:

$$\begin{aligned}
 & P \left\{ \sup_{1 \leq k \leq N} |R(\alpha_k) - R_{\text{emp}}(\alpha_k)| \hat{\theta}(R_{\text{emp}}(\alpha_k)) > \varepsilon \right\} \\
 & \leq \sum_{k=1}^N P \left\{ |R(\alpha_k) - R_{\text{emp}}(\alpha_k)| \hat{\theta}(R_{\text{emp}}(\alpha_k)) > \varepsilon \right\} \\
 & \leq (N-1) \sup_{1 \leq k \leq N} P \left\{ |R(\alpha_k) - R_{\text{emp}}(\alpha_k)| \hat{\theta}(R_{\text{emp}}(\alpha_k)) > \varepsilon \right\} \leq (N-1)P_\varepsilon,
 \end{aligned} \tag{4.13}$$

where P_ε is the probability that a function with probability of error larger than ε has zero empirical risk (zero frequency of error). (Note that in (4.13) we have coefficient $N-1$ rather than N since at least one of the probabilities in the sum is equal to zero.)

This probability can be easily bounded:

$$P_\varepsilon \leq (1-\varepsilon)^\ell.$$

Substituting the bound of P_ε into (4.13), we obtain

$$P \left\{ \sup_{1 \leq k \leq N} |R(\alpha_k) - R_{\text{emp}}(\alpha_k)| \hat{\theta}(R_{\text{emp}}(\alpha_k)) > \varepsilon \right\} \leq (N-1)(1-\varepsilon)^\ell. \tag{4.14}$$

As was done in the last section we rewrite this inequality in equivalent form. To do this let us consider the equality for arbitrary $0 \leq \eta \leq 1$:

$$(N-1)(1-\varepsilon)^\ell = \eta, \quad N > 1$$

and solve it with respect to ε :

$$\varepsilon = 1 - \exp \left\{ -\frac{\ln(N-1) - \ln \eta}{\ell} \right\} \leq \frac{\ln(N-1) - \ln \eta}{\ell}, \quad N > 1$$

Now we can rewrite inequality (4.14) in the following equivalent form:

With probability $1-\eta$ simultaneously all functions $Q(z, \alpha_k^)$ from a given finite set of functions that have empirical risk equal to zero satisfy the inequality*

$$R(\alpha_k^*) \leq \frac{\ln(N-1) - \ln \eta}{\ell}, \quad N > 1. \tag{4.15}$$

This bound is tight. It is achieved when the set of functions $Q(z, \alpha_k)$, $k = 1, 2, \dots, N$, contains one function with value of risk equal to zero and the remaining $N-1$ functions form statistically independent events $A_k = \{z : Q(z, \alpha_k) > 0\}$ (with respect to probability measure $F(z)$) and have the same value ε of error probability.

For this optimistic case the minimal value of the risk equals zero. Therefore with probability $1 - \eta$ one can assert that the difference between the value of the guaranteed risk and the value of the best possible risk has the bound

$$\Delta(\alpha_k^*) < \frac{\ln(N-1) - \ln \eta}{\ell} \quad (4.16)$$

Thus the bounds (4.15) and (4.16) give complete information about the generalization ability of the principle of empirical risk minimization for the optimistic case of the simplest model.

Note that in the optimistic case we have obtained significantly better bounds for the generalization ability than in the pessimistic case (the bounds are proportional to $1/\ell$ instead of $1/\sqrt{\ell}$).

4.3 BOUNDS FOR THE SIMPLEST MODEL: GENERAL CASE

The bounds for the Simplest Model that combine in one formula both the bound for the pessimistic case and the bound for the optimistic case and (what is more important) consider the intermediate cases based on the *multiplicative Chernoff* inequalities: For ℓ random independent trials in the Bernoulli scheme the inequalities

$$P \left\{ \frac{p - \nu_\ell}{\sqrt{p}} > \varepsilon \right\} < \exp \left\{ \frac{-\varepsilon^2 \ell}{2} \right\}, \quad (4.17)$$

$$P \left\{ \frac{\nu_\ell - p}{\sqrt{p}} > \varepsilon \right\} < \exp \left\{ \frac{-\varepsilon^2 \ell}{3} \right\} \quad (4.18)$$

are valid.[†]

Now consider the finite set of indicator functions $Q(z, \alpha_k)$, $k = 1, 2, \dots, N$. Using the inequality (4.17) one can derive (as was done in previous sections) that the following inequality is valid:

$$P \left\{ \sup_{1 \leq k \leq N} \frac{R(\alpha_k) - R_{\text{emp}}(\alpha_k)}{\sqrt{R(\alpha_k)}} > \varepsilon \right\} < N \exp \left\{ \frac{-\varepsilon^2 \ell}{2} \right\}. \quad (4.19)$$

Let us rewrite this inequality in the equivalent form.

[†]These bounds usually are given in the following equivalent form:

$$\begin{aligned} P\{\nu_\ell < (1 - \gamma)p\} &< \exp \left\{ -\frac{\gamma^2 p \ell}{2} \right\}, \\ P\{\nu_\ell > (1 + \gamma)p\} &< \exp \left\{ -\frac{\gamma^2 p \ell}{3} \right\}. \end{aligned}$$

With probability $1 - \eta$ simultaneously for all N functions in the set $Q(z, \alpha_k)$, $k = 1, 2, \dots, N$, the inequality

$$R(\alpha_k) < R_{\text{emp}}(\alpha_k) + \frac{\ln N - \ln \eta}{\ell} \left(1 + \sqrt{1 + 2 \frac{R_{\text{emp}}(\alpha_k)\ell}{\ln N - \ln \eta}} \right) \quad (4.20)$$

holds true.

To obtain these bounds, one has to equate the right-hand side of inequality (4.10) to some positive value $0 < \eta \leq 1$

$$N \exp \left\{ \frac{-\varepsilon^2 \ell}{2} \right\} = \eta$$

and solve this equation with respect to ε

$$\varepsilon = \sqrt{2 \frac{\ln N - \ln \eta}{\ell}}$$

Then using this ε one can obtain (4.20) as the solution of the inequality

$$\frac{R(\alpha_k) - R_{\text{emp}}(\alpha_k)}{\sqrt{R(\alpha_k)}} \leq \varepsilon.$$

Since with probability at least $1 - \eta$ inequality (4.20) is true for all N functions in a given set, it is true in particular for the function $Q(z, \alpha_{k(\ell)})$ which minimizes the empirical risk functional.

For this function with probability $1 - \eta$ the bound

$$R(\alpha_{k(\ell)}) < R_{\text{emp}}(\alpha_{k(\ell)}) + \frac{\ln N - \ln \eta}{\ell} \left(1 + \sqrt{1 + 2 \frac{R_{\text{emp}}(\alpha_{k(\ell)})\ell}{\ln N - \ln \eta}} \right) \quad (4.21)$$

holds true.

To estimate how close the risk $R(\alpha_{k(\ell)})$ is to the minimal risk for this set of functions let us define a lower bound on the risk for the function $Q(z, \alpha_{k(0)})$ which minimizes the expected risk. To do this we rewrite for this function the additive Chernoff bound (4.11) in the following equivalent form: With probability at least $1 - \eta$ the inequality

$$R(\alpha_{k(0)}) > R_{\text{emp}}(\alpha_{k(0)}) - \sqrt{\frac{-\ln \eta}{2\ell}}$$

holds true.

Using this bound and the bound (4.21) we obtain that with probability $1 - 2\eta$ the inequality

$$\begin{aligned}\Delta(\alpha_{k(\ell)}) &= R(\alpha_{k(\ell)}) - R(\alpha_{k(0)}) \\ &< \sqrt{\frac{-\ln \eta}{2\ell}} + \frac{\ln N - \ln \eta}{\ell} \left(1 + \sqrt{1 + 2 \frac{R_{\text{emp}}(\alpha_{k(\ell)})\ell}{\ln N - \ln \eta}} \right)\end{aligned}\quad (4.22)$$

is valid.

The inequalities (4.21), (4.22) describe the generalization ability of the method of empirical risk minimization for the Simplest Model.

Note that when the empirical risk equals zero the bound (4.21) differ from the bound (4.15) (derived for the optimistic case) only by the factor of 2. When the value of the empirical risk is close to 1/2 and ℓ is rather large, the bound (4.22) is close to the bound (4.9) derived for the pessimistic case.

The next sections of this chapter are devoted to deriving the bounds on the generalization ability for an infinite set of indicator functions $Q(z, a), a \in A$. First we derive the bounds for pessimistic case and then, using them we derive the bounds for the general case. The bounds for infinite sets of functions have the same form as the bounds for finite sets of functions. However, instead of a logarithm of the number of functions in the set we shall use another measure of the capacity of a set of functions.

4.4 THE BASIC INEQUALITIES: PESSIMISTIC CASE

Now let a set of indicator functions $Q(z, a), a \in A$, contain an infinite number of elements. As before, our goal is to estimate the rate of uniform convergence of the frequencies $R_{\text{emp}}(\alpha)$ to their probabilities $R(\alpha)$.

Let

$$z_1, \dots, z_\ell \quad (4.23)$$

be a random independent observation of size ℓ .

Let

$$N^A(z_1, \dots, z_\ell) \leq 2^\ell$$

be the number of different separations of the sample (4.23) by a given set of functions. Assume that $N^A(z_1, \dots, z_\ell)$ is measurable with respect to the probability measure $F(z_1, \dots, z_\ell)$. Therefore the expectation $EN^A(z_1, \dots, z_\ell)$ exists.

In the last section of Chapter 3 we introduced the concept of annealed entropy of a set of indicator functions on a sample of size ℓ :

$$H_{\text{ann}}^A(\ell) = \ln EN^A(z_1, \dots, z_\ell). \quad (4.24)$$

Using this concept we formulate the basic theorem of the theory of the rate of uniform convergence, which we will prove in the next section.

Theorem 4.1. *The inequality*

$$P \left\{ \sup_{\alpha \in \Lambda} |R(\alpha) - R_{\text{emp}}(\alpha)| > \varepsilon \right\} < 4 \exp \left\{ \left(\frac{H_{\text{ann}}^{\Lambda}(2\ell)}{\ell} - \left(\varepsilon - \frac{1}{\ell} \right)^2 \right) \ell \right\} \quad (4.25)$$

holds true.

Corollary. *For the existence of nontrivial exponential bounds on uniform convergence,*

$$\lim_{\ell \rightarrow \infty} \frac{H_{\text{ann}}^{\Lambda}(\ell)}{\ell} = 0$$

is sufficient.

In Chapter 3 we called this equation *the second milestone in the learning theory*.

In the next section we prove this theorem; however, before that we rewrite the inequality (4.25) in the equivalent form.

With probability at least $1 - \eta$ simultaneously for all functions in the set $Q(z, a)$, $a \in A$, the inequality

$$R(\alpha) \leq R_{\text{emp}}(\alpha) + \sqrt{\frac{H_{\text{ann}}^{\Lambda}(2\ell) - \ln \eta/4}{\ell}} + \frac{1}{\ell}$$

holds true.

In particular this inequality holds for the function $Q(z, \alpha_\ell)$, which minimizes the empirical risk functional. Thus with probability at least $1 - \eta$ the inequality

$$R(\alpha_\ell) \leq R_{\text{emp}}(\alpha_\ell) + \sqrt{\frac{H_{\text{ann}}^{\Lambda}(\ell) - \ln \eta/4}{\ell}} + \frac{1}{\ell} \quad (4.26)$$

holds true. As was shown in Section 4.1 for the function $Q(z, \alpha_0)$ which minimizes the expected risk, with probability $1 - \eta$ inequality

$$R(\alpha_0) > R_{\text{emp}}(\alpha_0) - \sqrt{\frac{-\ln \eta}{2\ell}}$$

is valid. From these two inequalities we obtain that with probability at least $1 - 2\eta$ the inequality

$$\Delta(\alpha_\ell) < R(\alpha_\ell) - R(\alpha_0) = \sqrt{\frac{H_{\text{ann}}^{\Lambda}(2\ell) - \ln \eta/4}{\ell}} - \sqrt{\frac{-\ln \eta}{2\ell}} + \frac{1}{\ell} \quad (4.27)$$

holds true.

The inequalities (4.26) and (4.27) have the same form as inequalities (4.9) and (4.12) in the simplest model. The only difference is that here we use a different concept of capacity, namely, annealed entropy $H_{\text{ann}}^{\Lambda}(\ell)$. These inequalities form one (from the two) pair of basic inequalities in the theory of the generalization ability of learning machines. The second pair of basic inequalities will be derived in Section 4.6, but first we prove Theorem 4.1.

4.5 PROOF OF THEOREM 4.1

The proof of Theorem 4.1 is based on the following lemma.

4.5.1 The Basic Lemma

Let us consider a space of random independent observations of size 2ℓ :

$$Z^{2\ell} = z_1, \dots, z_\ell, z_{\ell+1}, \dots, z_{2\ell}.$$

For any function in the set $Q(z, a)$, $a \in A$, we determine the frequency

$$\nu(\alpha, Z_1) = \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha)$$

on the first part of a sample

$$Z_1 = z_1, \dots, z_\ell$$

and determine the frequency

$$\nu(\alpha, Z_2) = \frac{1}{\ell} \sum_{i=\ell+1}^{2\ell} Q(z_i, \alpha)$$

on the second part of a sample

$$Z_2 = z_{\ell+1}, \dots, z_{2\ell}.$$

Let us denote by $Z_1(\ell)$ and $Z_2(\ell)$ two spaces of half-samples of length 4. Consider the random variables

$$\rho^{\Lambda}(\alpha, Z^{2\ell}) = \left| \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) - \frac{1}{\ell} \sum_{i=\ell+1}^{2\ell} Q(z_i, \alpha) \right|,$$

$$\rho^{\Lambda}(Z^{2\ell}) = \sup_{\alpha \in \Lambda} \left| \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) - \frac{1}{\ell} \sum_{i=\ell+1}^{2\ell} Q(z_i, \alpha) \right|,$$

and consider the random variable

$$\begin{aligned}\pi^\Lambda(\alpha, Z_1) &= \left| \int Q(\alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \right|, \\ \pi^\Lambda(Z_1) &= \sup_{\alpha \in \Lambda} \left| \int Q(\alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \right|.\end{aligned}$$

We assume that functions $\pi^\Lambda(Z_1)$ and $\rho^\Lambda(Z^{2\ell})$ are measurable with respect to probability measures defined by $F(z)$. So $\rho^\Lambda(Z^{2\ell})$ and $\pi^\Lambda(Z_1)$ are random variables.

Basic Lemma. *The distribution of the random variable $\pi^\Lambda(Z_1)$ is connected with the distribution of the random variable $\rho^\Lambda(Z^{2\ell})$ by the inequality*

$$P \left\{ \pi^\Lambda(Z_1) > \varepsilon \right\} < 2P \left\{ \rho^\Lambda(Z^{2\ell}) > \varepsilon - \frac{1}{\ell} \right\}. \quad (4.28)$$

Therefore according to the basic lemma to estimate the rate of convergence to zero of the random variable $\pi^\Lambda(Z_1)$ one can estimate the rate of convergence to zero of the random variable $\rho^\Lambda(Z^{2\ell})$.

Below we first prove this lemma, then describe the idea of how to estimate the rate of convergence of the random variable $\rho^\Lambda(Z^{2\ell})$, and at the end give the formal proof of the theorem.

4.5.2 Proof of Basic Lemma

By definition,

$$P \left\{ \rho^\Lambda(Z^{2\ell}) > \varepsilon - \frac{1}{\ell} \right\} = \int_{Z(2\ell)} \theta \left[\rho^\Lambda(Z^{2\ell}) - \varepsilon + \frac{1}{\ell} \right] dF(Z^{2\ell}).$$

Taking into account that the space $Z(2\ell)$ of samples of size 2ℓ is the direct product of $Z_1(\ell)$ and $Z_2(\ell)$ of half samples of size ℓ , by Fubini's theorem we have

$$P \left\{ \rho^\Lambda(Z^{2\ell}) > \varepsilon - \frac{1}{\ell} \right\} = \int_{Z_1(\ell)} dF(Z_1) \int_{Z_2(\ell)} \theta \left[\rho^\Lambda(Z^{2\ell}) - \varepsilon + \frac{1}{\ell} \right] dF(Z_2)$$

(in the inner integral the first half of the sample is fixed). Denote by Q the following event in the space $Z_1(\ell)$:

$$Q = \left\{ Z_1 : \pi^\Lambda(Z_1) > c \right\}.$$

Reducing the domain of integration we obtain

$$P \left\{ \rho^\Lambda(Z^{2\ell}) > \varepsilon - \frac{1}{\ell} \right\} \geq \int_Q dF(Z_1) \int_{Z_2(\ell)} \theta \left[\rho^\Lambda(Z^{2\ell}) - \varepsilon + \frac{1}{\ell} \right] dF(Z_2). \quad (4.29)$$

We now bound the inner integral on the right-hand side of the inequality which we denote by I . Recall that here the sample Z_1 is fixed and is such that

$$\pi^\Lambda(Z_1) > \varepsilon.$$

Consequently there exists an $a^* \in A$ such that

$$|P(a^*) - \nu(a^*, Z_1)| > \varepsilon,$$

where we denote

$$P(a^*) = \int Q(z, a^*) dF(z).$$

Then

$$\begin{aligned} I &= \int_{Z_2(\ell)} \theta \left[\sup_{\alpha \in \Lambda} \rho(\alpha, Z^{2\ell}) - \varepsilon + \bar{\varepsilon} \right] dF(Z_2) \\ &\geq \int_{Z_2(\ell)} \theta \left[\rho(a^*, Z^{2\ell}) - \varepsilon + \frac{1}{\ell} \right] dF(Z_2). \end{aligned}$$

Now let

$$\nu(a^*, Z_1) < P(a^*) - \varepsilon$$

(the case $\nu(a^*, Z_1) \geq P(a^*) - \varepsilon$ is dealt with completely analogously). Then in order for the condition

$$|\nu(a^*, Z_1) - \nu(a^*, Z_2)| > \varepsilon - \frac{1}{\ell}$$

to be satisfied, it is sufficient that the relation

$$\nu(a^*, Z_2) \geq P(a^*) - \frac{1}{\ell}$$

holds, from which we obtain

$$\begin{aligned} I &\geq \int_{Z_2} \theta \left[\nu(a^*, Z_2) \geq P(a^*) - \frac{1}{\ell} \right] dF(Z_2) = P \{ \nu(a^*, Z_2) \geq P(a^*) \} \\ &= \sum_{k/\ell \geq P(a^*) - 1/\ell} C_\ell^k [P(a^*)]^k [1 - P(a^*)]^{\ell-k}. \end{aligned}$$

The last sum exceeds $\frac{1}{2}$. Therefore returning to (4.29) we obtain

$$P \left\{ \rho^\Lambda(Z^{2\ell}) > \varepsilon - \frac{1}{\ell} \right\} > \frac{1}{2} \int_Q dF(Z_1) = \frac{1}{2} P \left\{ \pi^\Lambda(Z_1) > \varepsilon \right\}.$$

The lemma is proved.

4.5.3 The Idea of Proving Theorem 4.1

Below we give an idea of proving Theorem 4.1. The formal proof will be given in the next section.

Denote

$$\varepsilon_* = \varepsilon - 1/\ell.$$

Suppose the sample

$$z_1, \dots, z_{2\ell}$$

is split *randomly* into two half-samples

$$z_1, \dots, z_\ell \quad \text{and} \quad z_{\ell+1}, \dots, z_{2\ell}.$$

For any fixed sample of size 2ℓ , any function $Q(z, a^*)$ and any two randomly chosen half-samples the classical inequality

$$P \left\{ \left| \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha^*) - \frac{1}{\ell} \sum_{i=\ell+1}^{2\ell} Q(z_i, \alpha^*) \right| > \varepsilon_* \right\} < 2 \exp\{-\varepsilon_*^2 \ell\}$$

holds true.

To estimate the probability that the largest deviation over a given set of functions exceeds ε_* we note the following. For any fixed sample $Z^{2\ell}$ there exists only a finite set of distinguishable functions $Q(z, a^*) \in A^* = A^*(z_1, \dots, z_{2\ell})$ of cardinality $N^A(z_1, \dots, z_{2\ell})$. Therefore we can use the same reasoning that we used for the Simplest Model

$$\begin{aligned} & P \left\{ \sup_{\alpha \in \Lambda} \left| \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) - \frac{1}{\ell} \sum_{i=\ell+1}^{2\ell} Q(z_i, \alpha) \right| > \varepsilon_* \mid z_1, \dots, z_{2\ell} \right\} \\ &= P \left\{ \sup_{\alpha^* \in \Lambda^*} \left| \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha^*) - \frac{1}{\ell} \sum_{i=\ell+1}^{2\ell} Q(z_i, \alpha^*) \right| > \varepsilon_* \mid z_1, \dots, z_{2\ell} \right\} \\ &\leq \sum_{\alpha^* \in \Lambda^*} P \left\{ \left| \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha^*) - \frac{1}{\ell} \sum_{i=\ell+1}^{2\ell} Q(z_i, \alpha^*) \right| > \varepsilon_* \mid z_1, \dots, z_{2\ell} \right\} \\ &\leq 2N^A(z_1, \dots, z_{2\ell}) \exp\{-\varepsilon_*^2 \ell\}. \end{aligned}$$

To get the bounds for a random sample of size 2ℓ it is sufficient to take expectation with respect to probability measure on the sample space

$$\begin{aligned} & P \left\{ \sup_{\alpha \in \Lambda} \left| \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) - \frac{1}{\ell} \sum_{i=\ell+1}^{2\ell} Q(z_i, \alpha) \right| > \varepsilon_* \right\} \\ &= EP \left\{ \sup_{\alpha \in \Lambda} \left| \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) - \frac{1}{\ell} \sum_{i=\ell+1}^{2\ell} Q(z_i, \alpha) \right| > \varepsilon_* \mid z_1, \dots, z_{2\ell} \right\} \\ &< EN^A(z_1, \dots, z_{2\ell}) \exp\{-\varepsilon_*^2 \ell\} = \exp \left\{ \left(\frac{H_{\text{ann}}^A(2\ell)}{\ell} - \varepsilon_*^2 \right) \ell \right\}. \end{aligned}$$

Combining this bound with the statement (4.28) of basic lemma proves the theorem.

4.5.4 Proof of Theorem 4.1

Now we proceed with the formal proof of this theorem. Let us denote by $Z(2\ell)$ the space of samples of size 2ℓ and by $Z^{2\ell} = (z_1, \dots, z_{2\ell})$ the specific sample. In view of the inequality (4.28) it is sufficient to bound the quantity

$$P \{ p''(Z^{2\ell}) > \varepsilon_* \} = \int_{Z(2\ell)} \theta \left[\rho^\Lambda(Z^{2\ell}) - \varepsilon_* \right] dF(Z^{2\ell}).$$

Consider the mapping of the space $Z(2\ell)$ into itself obtaining by a permutation T_i of the elements of sequence $Z^{2\ell}$. There are $(2\ell)!$ different permutations of the sample of size 2ℓ .

In view of symmetry of the definition of the measure, the equality

$$\int_{Z(2\ell)} f(Z^{2\ell}) dF(Z^{2\ell}) = \int_{Z(2\ell)} f(T_i Z^{2\ell}) dF(Z^{2\ell}), \quad i = 1, 2, \dots, (2\ell)!$$

holds for any integrable function $f(Z^{2\ell})$. Therefore

$$P \left\{ \rho^\Lambda(Z^{2\ell}) > \varepsilon_* \right\} = \int_{Z(2\ell)} \frac{\sum_{i=1}^{(2\ell)!} 8 [\rho(T_i Z^{2\ell}) - \varepsilon_*]}{(2\ell)!} dF(Z^{2\ell}). \quad (4.30)$$

Observe that

$$\begin{aligned} \theta \left[\rho^\Lambda(Z^{2\ell}) - \varepsilon_* \right] &= \theta \left[\sup_{\alpha \in \Lambda} \left| \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) - \frac{1}{\ell} \sum_{i=\ell+1}^{2\ell} Q(z_i, \alpha) \right| - \varepsilon_* \right] \\ &= \sup_{\alpha \in \Lambda} 8 \left[\left| \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) - \frac{1}{\ell} \sum_{i=\ell+1}^{2\ell} Q(z_i, \alpha) \right| - \varepsilon_* \right]. \end{aligned}$$

Clearly if two functions $Q(z, \alpha_1)$ and $Q(z, \alpha_2)$ are nondistinguishable on the sample $z_1, \dots, z_{2\ell}$, then

$$\rho^\Lambda(T_i Z^{2\ell}, \alpha_1) = \rho^\Lambda(T_i Z^{2\ell}, \alpha_2)$$

is true for any permutation T_i . In other words, if two functions are equivalent with respect to the sample $z_1, \dots, z_{2\ell}$, then deviations in frequencies for these two functions are the same for all permutations T_i . Therefore, for each class of equivalent functions one can choose only one function $Q(z, \alpha^*)$ which forms a finite set of functions $Q(z, \alpha^*) \in A^* \subset \Lambda$ such that

$$\sup_{\alpha \in \Lambda} \rho(T_i Z^{2\ell}, \alpha) = \sup_{\alpha^* \in \Lambda^*} \rho(T_i Z^{2\ell}, \alpha^*), \quad i = 1, \dots, (2\ell)!$$

The number of functions in the set A^* is finite and does not exceed $N^\Lambda(z_1, \dots, z_{2\ell})$. Replacing the sup operation by summation, we obtain

$$\begin{aligned} \sup_{\alpha \in \Lambda} \theta [\rho(T_i Z^{2\ell}, \alpha) - \varepsilon_*] &= \sup_{\alpha^* \in \Lambda^*} \theta [\rho(T_i Z^{2\ell}, \alpha^*) - \varepsilon_*] \\ &\leq \sum_{\alpha^* \in \Lambda^*} \theta [\rho(T_i Z^{2\ell}, \alpha^*) - \varepsilon_*] \end{aligned}$$

These relations allow us to bound the integrand in (4.30):

$$\begin{aligned} \frac{\sum_{i=1}^{(2\ell)!} \theta [\rho^\Lambda(T_i Z^{2\ell}) - \varepsilon_*]}{(2\ell)!} &= \frac{\sum_{i=1}^{(2\ell)!} \sup_{\alpha^* \in \Lambda^*} \theta [\rho(T_i Z^{2\ell}, \alpha^*) - \varepsilon_*]}{(2\ell)!} \\ &\leq \sum_{\alpha^* \in \Lambda^*} \frac{\sum_{i=1}^{(2\ell)!} \theta [\rho(T_i Z^{2\ell}, \alpha^*) - \varepsilon_*]}{(2\ell)!} \end{aligned}$$

Note that the summand on the right-hand side of the last inequality is the ratio of the number of orderings in a sample (of fixed composition) such that

$$\left| \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha^*) - \frac{1}{\ell} \sum_{i=\ell+1}^{2\ell} Q(z_i, \alpha^*) \right| > \varepsilon_*$$

to the total number of permutations. It is easy to see that this value is equal to

$$\Gamma = \sum_k \frac{C_m^k C_{2\ell-m}^{\ell-k}}{C_{2\ell}^{\ell}},$$

$$\left\{ k : \left| \frac{k}{\ell} - \frac{m-k}{\ell} \right| > \varepsilon_* \right\},$$

where m is the number of elements z_j in the sample $z_1, \dots, z_{2\ell}$ for which $Q(z_j, \alpha^*) = 1$.

In Section 4.13 we shall obtain the following bound for Γ

$$\Gamma < 2 \exp \left\{ -\varepsilon_*^2 \ell \right\}.$$

Thus

$$\begin{aligned} \sum_{\alpha^* \in \Lambda^*} \frac{\sum_{i=1}^{(2\ell)!} \theta [\rho(T_i Z^{2\ell}, \alpha^*) - \varepsilon_*]}{(2\ell)!} &< 2 \sum_{\alpha^* \in \Lambda^*} \exp \left\{ -\varepsilon_*^2 \ell \right\} \\ &= 2 N^\Lambda(z_1, \dots, z_{2\ell}) \exp \left\{ -\varepsilon_*^2 \ell \right\}. \end{aligned}$$

Substituting this bound into integral (4.30), we obtain

$$\begin{aligned} P \left\{ \rho^\Lambda(Z^{2\ell}) > \varepsilon_* \right\} &< 2EN^\Lambda(z_1, \dots, z_{2\ell}) \exp \left\{ -\varepsilon_*^2 \ell \right\} \\ &= 2 \exp \left\{ \left(\frac{H_{\text{ann}}^\Lambda(2\ell)}{\ell} - \varepsilon_*^2 \right) \ell \right\}, \end{aligned}$$

from which, in view of the basic lemma, we obtain

$$P \{ \pi^\Lambda(Z^{2\ell}) > \varepsilon \} < 4 \exp \left\{ \left(\frac{H_{\text{ann}}^\Lambda(2\ell)}{\ell} - \varepsilon_*^2 \right) \ell \right\}$$

Recalling that we denote $\varepsilon_* = \varepsilon - 1/\ell$ we obtain the desired bound. The theorem is proved.

4.6 BASIC INEQUALITIES: GENERAL CASE

In this section we discuss the theorem about the rate of relative uniform convergence of frequencies to their probabilities. (Proof of this theorem is given in the next section.)

Theorem 4.2. *For any ℓ the inequality*

$$P \left\{ \sup_{\alpha \in \Lambda} \frac{R(\alpha) - R_{\text{emp}}(\alpha)}{\sqrt{R(\alpha)}} > \varepsilon \right\} < 4 \exp \left\{ \left(\frac{H_{\text{ann}}^\Lambda(2\ell)}{\ell} - \frac{\varepsilon^2}{4} \right) \ell \right\} \quad (4.31)$$

holds true.

Corollary. *For the existence of nontrivial exponential bounds on uniform relative convergence it is sufficient that*

$$\lim_{\ell \rightarrow \infty} \frac{H_{\text{ann}}^\Lambda(\ell)}{\ell} = 0. \quad (4.32)$$

Let us rewrite inequality (4.31) in the equivalent form. As before we equate the right-hand side of inequality (4.31) to a positive value η ($0 < \eta \leq 1$)

$$4 \exp \left\{ \left(\frac{H_{\text{ann}}^\Lambda(2\ell)}{\ell} - \frac{\varepsilon^2}{4} \right) \ell \right\} = \eta$$

and solve this equation with respect to ε^2 . The solution

$$\varepsilon(\ell) = \varepsilon^2 = 4 \frac{H_{\text{ann}}^\Lambda(2\ell) - \ln \eta/4}{\ell}$$

is used to solve inequality

$$\sup_{\alpha \in \Lambda} \frac{R(\alpha) - R_{\text{emp}}(\alpha)}{\sqrt{R(\alpha)}} \leq \mathcal{E}(\ell).$$

As a result we obtain that *with probability at least $1 - \eta$ simultaneously for all functions in the set of indicator functions $Q(z, \alpha)$, $\alpha \in \Lambda$, the inequality*

$$R(\alpha) \leq R_{\text{emp}}(\alpha) + \frac{\mathcal{E}(\ell)}{2} \left(1 + \sqrt{1 + \frac{4R_{\text{emp}}(\alpha)}{\mathcal{E}(\ell)}} \right)$$

is valid.

Since with probability $1 - \eta$ this inequality holds for all functions of the set $Q(z, \alpha)$, $\alpha \in \Lambda$, it holds in particular for the function $Q(z, \alpha_\ell)$ which minimizes the empirical risk functional. For this function with probability $1 - \eta$ the bound

$$R(\alpha_\ell) \leq R_{\text{emp}}(\alpha_\ell) + \frac{\mathcal{E}(\ell)}{2} \left(1 + \sqrt{1 + \frac{4R_{\text{emp}}(\alpha_\ell)}{\mathcal{E}(\ell)}} \right) \quad (4.33)$$

holds true.

Taking into account that for the function $Q(z, \alpha_0)$ which minimizes the *expected risk* in the set of functions $Q(z, \alpha)$, $\alpha \in A$, the additive Chernoff inequality (4.4) holds true, one can assert that with probability at least $1 - \eta$ the inequality

$$R(\alpha_0) > R_{\text{emp}}(\alpha_0) - \sqrt{\frac{-\ln \eta}{2\ell}}$$

is valid.

Note that

$$R_{\text{emp}}(\alpha_0) \geq R_{\text{emp}}(\alpha_\ell) \quad (4.34)$$

From (4.33), the lower bound for $R(\alpha_0)$, and (4.34) we deduce that with probability at least $1 - 277$ the inequality

$$\begin{aligned} \Delta(\alpha_\ell) &= R(\alpha_\ell) - R(\alpha_0) \\ &< \sqrt{\frac{-\ln \eta}{2\ell}} + \frac{\mathcal{E}(\ell)}{2} \left(1 + \sqrt{1 + \frac{4R_{\text{emp}}(\alpha_\ell)}{\mathcal{E}(\ell)}} \right) \end{aligned} \quad (4.35)$$

is valid.

Thus, the bounds (4.33) and (4.35) describe the generalization ability of algorithms that minimize empirical risk: Bound (4.33) evaluates the risk for the chosen function, and bound (4.35) evaluates how close this risk is to the smallest possible risk for a given set of functions.

These two bounds are the basic bounds for the generalization ability of algorithms that minimize empirical risk in the problem of pattern recognition. They have exactly the same form as the bound for generalization ability in the simplest model. The only difference is that here we use a more sophisticated concept of capacity than in the simplest model.

4.7 PROOF OF THEOREM 4.2

In this section we prove a more general version of Theorem 4.2 to be used in the next chapter.

Theorem 4.2*. *For any $1 < p \leq 2$ the inequality*

$$P \left\{ \sup_{\alpha \in \Lambda} \frac{R(\alpha) - R_{\text{emp}}(\alpha)}{\sqrt[\ell]{R(\alpha)}} > \varepsilon \right\} < 4 \exp \left\{ \left(\frac{H_{\text{ann}}^{\Lambda}(2\ell)}{\ell^{2-2/p}} - \frac{\varepsilon^2}{2^{1+2/p}} \right) \ell^{2-2/p} \right\} \quad (4.35a)$$

holds true.

Consider two events constructed from a random and independent sample of size 2ℓ :

$$\begin{aligned} Q_1 &= \left\{ z : \sup_{\alpha \in \Lambda} \frac{P(A_\alpha) - \nu_1(A_\alpha)}{\sqrt[\ell]{P(A_\alpha)}} > \varepsilon \right\}, \\ Q_2 &= \left\{ z : \sup_{\alpha \in \Lambda} \frac{\nu_2(A_\alpha) - \nu_1(A_\alpha)}{\sqrt[\ell]{\nu(A_\alpha) + \frac{1}{2\ell}}} > \varepsilon \right\}, \end{aligned}$$

where A_α is the event

$$A_\alpha = \{z : Q(z, \alpha) = 1\},$$

$P(A_\alpha)$ is probability of event A_α ,

$$P(A_\alpha) = \int Q(z, \alpha) dF(z),$$

$\nu_1(A_\alpha)$ is the frequency of event A_α computed from the first half-sample z_1, \dots, z_ℓ of the sample $z_1, \dots, z_\ell, z_{\ell+1}, \dots, z_{2\ell}$

$$\nu_1(A_\alpha) = \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha),$$

and $\nu_2(A_\alpha)$ is the frequency of event A_α computed from the second half-

sample $z_{\ell+1}, \dots, z_{2\ell}$

$$\nu_2(A_\alpha) = \frac{1}{\ell} \sum_{i=\ell+1}^{2\ell} Q(z_i, \alpha).$$

Denote

$$\nu(A_\alpha) = \frac{\nu_1(A_\alpha) + \nu_2(A_\alpha)}{2}$$

Note that in case $\ell \leq \varepsilon^{-p/(p-1)}$ the assertion of the theorem is trivial (the right-hand side of inequality exceeds one). Accordingly we shall prove the theorem as follows: First we show that for $\ell > \varepsilon^{-p/(p-1)}$ the inequality

$$P(Q_1) < 4P(Q_2)$$

is valid, and then we bound $P(Q_2)$.

Thus we shall prove the lemma.

Lemma 4.1. *For $\ell > \varepsilon^{-p/(p-1)}$ the inequality*

$$P(Q_1) < 4P(Q_2)$$

is valid.

Proof Assume that event Q_1 has occurred. This means that there exists event A^* such that for the first half-sample the equality

$$P(A^*) - \nu_1(A^*) > \varepsilon \sqrt[p]{P(A^*)}$$

is fulfilled. Since $\nu_1(A^*) \geq 0$, this implies that

$$P(A^*) > \varepsilon^{p/(p-1)}.$$

Assume that for the second half-sample the frequency of event A^* exceeds the probability $P(A^*)$:

$$\nu_2(A^*) > P(A^*).$$

Recall now that $\ell > \varepsilon^{-p/(p-1)}$. Under these conditions, event Q_2 will definitely occur.

To show this we bound the quantity

$$\mu = \frac{\nu_2(A^*) - \nu_1(A^*)}{\sqrt[p]{\nu(A^*) + 1/2\ell}} \tag{4.36}$$

under the conditions

$$\nu_1(A^*) < P(A^*) - \varepsilon \sqrt[p]{P(A^*)},$$

$$\nu_2(A^*) > P(A^*),$$

$$P(A^*) > \varepsilon^{p/(p-1)}.$$

For this purpose we find the minimum of the function

$$T = \frac{x - y}{\sqrt[p]{x + y + c}}$$

in the domain $0 < a \leq x \leq 1$, $0 < y \leq b$, $c > 0$. We have for $p > 1$

$$\frac{\partial T}{\partial x} = \frac{1}{p} \frac{(p-1)x + (p+1)y + pc}{(x+y+c)^{(p+1)/p}} > 0,$$

$$\frac{\partial T}{\partial y} = -\frac{1}{p} \frac{(p+1)x + (p-1)y + pc}{(x+y+c)^{(p+1)/p}} < 0.$$

Consequently T attains its minimum in the admissible domain at the boundary points $x = a$ and $y = b$.

Therefore the quantity μ is bounded from below, if in (4.36) one replaces $\nu_1(A^*)$ by $P(A^*) - \varepsilon \sqrt[p]{P(A^*)}$ and $\nu_2(A^*)$ by $P(A^*)$. Thus

$$\mu \geq \frac{\varepsilon \sqrt[p]{2P(A^*)}}{\sqrt{2P(A^*) - \varepsilon \sqrt[p]{P(A^*)} + 1/\ell}}$$

Furthermore, since $P(A^*) > \varepsilon^{p/(p-1)}$ and $\ell > \varepsilon^{-p/(p-1)}$ we have that

$$\mu > \frac{\varepsilon \sqrt[p]{2P(A^*)}}{\sqrt[p]{2P(A^*) - \varepsilon^{(p+1)/p} + \varepsilon^{(p+1)/p}}} = \varepsilon.$$

Thus, if \mathcal{Q}_1 occurs and the conditions $\nu_2(A^*) > P(A^*)$ is satisfied, then \mathcal{Q}_2 occurs as well.

Observe that the second half-sample is chosen independently of the first one and that the frequency $\nu_2(A^*)$ exceeds $P(A^*)$ with probability at most $1/4$ if $\ell P(A^*) > 1$. Therefore, provided that \mathcal{Q}_1 is fulfilled, the event

$$\nu_2(A^*) > P(A^*)$$

occurs with probability exceeding $1/4$. Since under condition of the lemma $\ell P(A^*) > 1$ is valid we have

$$P(\mathcal{Q}_2) > \frac{1}{4}P(\mathcal{Q}_1).$$

The lemma is proved.

Lemma 4.2. For any $1 < p \leq 2$ and any $0 > \varepsilon^{-p/(p-1)}$ the bound

$$P(Q_2) < \exp \left\{ \left(\frac{H_{\text{ann}}^{\Lambda}(2\ell)}{\ell^{2-2/p}} - \frac{\varepsilon^2}{2^{1+2/p}} \right) \ell^{2-2/p} \right\}$$

is valid.

Proof. Denote by $R_A(Z^{2\ell})$ the quantity

$$R_A(Z^{2\ell}) = \frac{\nu_2(A) - \nu_1(A)}{\sqrt[p]{\nu(A) + 1/2\ell}}$$

then the estimated probability equals

$$P(Q_2) = \int_{Z(2\ell)} \theta \left[\sup_{A \in S} R_A(Z^{2\ell}) - \varepsilon \right] dF(Z^{2\ell})$$

Here the integration is carried out over the space of all possible samples of size 24.

Consider now all possible permutations T_i , $i = 1, 2, \dots, (2\ell)!$ of the sequence $z_1, \dots, z_{2\ell}$. For each such permutation the equality

$$\int_{Z(2\ell)} \theta \left[\sup_{A \in S} R_A(Z^{2\ell}) - \varepsilon \right] dF(Z^{2\ell}) = \int_{Z(2\ell)} \theta \left[\sup_{A \in S} R_A(T_i Z^{2\ell}) - \varepsilon \right] dF(Z^{2\ell})$$

is valid. Therefore the equality

$$\begin{aligned} P(Q_2) &= \int_{Z(2\ell)} \theta \left[\sup_{A \in S} R_A(Z^{2\ell}) - \varepsilon \right] dF(Z^{2\ell}) \\ &= \int_{Z(2\ell)} \frac{1}{(2\ell)!} \sum_{i=1}^{(2\ell)!} \theta \left[\sup_{A \in S} R_A(T_i Z^{2\ell}) - \varepsilon \right] dF(Z^{2\ell}) \end{aligned} \quad (4.37)$$

is valid.

Now consider the integrand. Since the sample $z_1, \dots, z_{2\ell}$ is fixed, instead of the system of events S one can consider a finite system of events S^* which contains one representative for each one of the equivalence classes. Thus the equality

$$\frac{1}{(2\ell)!} \sum_{i=1}^{(2\ell)!} \theta \left[\sup_{A \in S} R_A(T_i Z^{2\ell}) - \varepsilon \right] = \frac{1}{(2\ell)!} \sum_{i=1}^{(2\ell)!} \theta \left[\sup_{A \in S^*} R_A(T_i Z^{2\ell}) - \varepsilon \right]$$

is valid. Furthermore,

$$\begin{aligned} \frac{1}{(2\ell)!} \sum_{i=1}^{(2\ell)!} \theta \left[\sup_{A \in S^*} R_A(T_i Z^{2\ell}) - \varepsilon \right] &< \frac{1}{(2\ell)!} \sum_{i=1}^{(2\ell)!} \sum_{A \in S^*} \theta \left[R_A(T_i Z^{2\ell}) - \varepsilon \right] \\ &= \sum_{A \in S^*} \left\{ \frac{1}{(2\ell)!} \sum_{i=1}^{(2\ell)!} \theta \left[R_A(T_i Z^{2\ell}) - \varepsilon \right] \right\} \end{aligned}$$

The expression in the braces is the probability of greater than ε deviation of the frequencies in two half-samples for a fixed event A and a given composition of a complete sample. This probability equals

$$\Gamma = \sum_k \frac{C_m^k C_{2\ell-m}^{\ell-k}}{C_{2\ell}^\ell}$$

where m is number of occurrences of event A in a complete sample, and k is number of occurrences of the event in the first half sample; k runs over the values

$$\begin{aligned} \max(0, m - \ell) \leq k \leq \min(m, \ell) \\ \frac{k}{\ell} - \frac{m - k}{\ell} > \varepsilon. \end{aligned}$$

Denote by ε^* the quantity

$$\sqrt[p]{\frac{m+1}{2\ell}} \varepsilon = \varepsilon^*.$$

Using this notation the constraints become

$$\begin{aligned} \max(0, m - \ell) \leq k \leq \min(m, \ell) \\ \frac{k}{\ell} - \frac{m - k}{\ell} > \varepsilon^*. \end{aligned} \tag{4.38}$$

In Section 4.13 the following bound on the quantity Γ under constraints (4.38) is obtained:

$$\Gamma < \exp \left\{ - \frac{(\ell + 1)(\varepsilon^*)^2 \ell^2}{(m + 1)(2\ell - m + 1)} \right\}. \tag{4.39}$$

Expressing (4.39) in terms of ε we obtain

$$\Gamma < \exp \left\{ - \frac{(\ell + 1)\varepsilon^2 \ell^2}{(m + 1)(2\ell - m + 1)} \left(\frac{m + 1}{2\ell} \right)^{2/p} \right\}$$

The right-hand side of this inequality reaches its maximum at $m = 0$. Thus

$$\Gamma < \exp \left\{ -\frac{\varepsilon^2}{2^{1+2/p}} \ell^{2-2/p} \right\}. \quad (4.40)$$

Substituting (4.40) into the right-hand side of (4.37) and integrating we have

$$\begin{aligned} P(Q_2) &= \int_{Z(2\ell)} N^S(Z^{2\ell}) \exp \left\{ -\frac{\varepsilon}{2^{1+2/p}} \ell^{2-2/p} \right\} dF(Z^{2\ell}) \\ &< \exp \left\{ \left(\frac{H_{\ell^2-2/p}''}{\ell^{2-2/p}} - 2^{1+2/p} \right) \ell^{2-2/p} \right\} \end{aligned}$$

The lemma is thus proved.

The assertion of the Theorem 4.2* follows from the inequalities obtained in the Lemma 4.1 and Lemma 4.2.

4.8 MAIN NONCONSTRUCTIVE BOUNDS

Thus, in the previous sections we obtained the basic bounds describing the generalization ability of learning machines that minimize the empirical risk functional:

1. With probability $1 - \eta$ any of the bounds

$$\begin{aligned} R(\alpha_\ell) &\leq R_{\text{emp}}(\alpha_\ell) + \sqrt{\mathcal{E}(\ell)} + \frac{1}{\ell}, \\ R(\alpha_\ell) &< R_{\text{emp}}(\alpha_\ell) + \frac{\mathcal{E}(\ell)}{2} \left(1 + \sqrt{\left(1 + \frac{4R_{\text{emp}}(\alpha_\ell)}{\mathcal{E}(\ell)} \right)} \right) \end{aligned} \quad (4.41)$$

hold true.

2. With probability $1 - 27$ any of the bounds

$$\begin{aligned} \Delta(\alpha_\ell) &< \sqrt{\mathcal{E}(\ell)} + \sqrt{\frac{-\ln \eta}{2\ell}} + \frac{1}{\ell}, \\ \Delta(\alpha_\ell) &< \frac{\mathcal{E}(\ell)}{2} \left(1 + \sqrt{\left(1 + \frac{4R_{\text{emp}}(\alpha_\ell)}{\mathcal{E}(\ell)} \right)} \right) + \sqrt{\frac{-\ln \eta}{2\ell}} \end{aligned} \quad (4.42)$$

hold true.

In these bounds we denote

$$\mathcal{E}(\ell) = 4 \frac{H_{\text{ann}}^\Lambda(2\ell) - \ln \eta/4}{\ell} \quad (4.43)$$

These bounds, however, are valid for a specific problem that is defined by probability measure $F(z)$ since the term $\mathcal{E}(\ell)$ that comes in the inequalities depends on the annealed entropy $H_{\text{ann}}^{\Lambda}(\ell)$ constructed on the basis of the unknown probability measure $F(z)$.

To make the bounds valid for any probability measure it is sufficient to use instead of the quantity $\mathcal{E}(\ell)$ the quantity

$$\mathcal{E}^*(\ell) = 4 \frac{G^{\Lambda}(2\ell) - \ln \eta/4}{\ell},$$

where the annealed entropy

$$H_{\text{ann}}^{\Lambda}(\ell) = E \ln N^{\Lambda}(z_1, \dots, z_\ell)$$

is replaced by the growth function

$$G^{\Lambda}(\ell) = \sup_{z_1, \dots, z_\ell} \ln N^{\Lambda}(z_1, \dots, z_\ell).$$

Since the growth function does not depend on the probability measure and is not less than the annealed entropy

$$H_{\text{ann}}^{\Lambda}(\ell) \leq G^{\Lambda}(\ell),$$

the bounds with $\mathcal{E}^*(\ell)$ (instead of $\mathcal{E}(\ell)$) are valid for any probability measure. These bounds are nontrivial if

$$\lim_{\ell \rightarrow \infty} \frac{G^{\Lambda}(\ell)}{\ell} = 0.$$

Note that the bounds with $\mathcal{E}^*(\ell)$ are upper bounds of the bounds with $\mathcal{E}(\ell)$.

Thus, we described the main bounds on the generalization ability of learning machines. These bounds, however, are nonconstructive since the theory does not suggest how to evaluate the growth function for a given set of functions.

Obtaining constructive bounds on the generalization ability of learning machines is based on the following remarkable property of the growth function.

4.9 VC DIMENSION

4.9.1 The Structure of the Growth Function

Theorem 43. *The growth function of a set of indicator functions $Q(z, a)$, $a \in A$ either (a) satisfies the equality*

$$G^{\Lambda}(\ell) = \ell \ln 2 \tag{4.44}$$

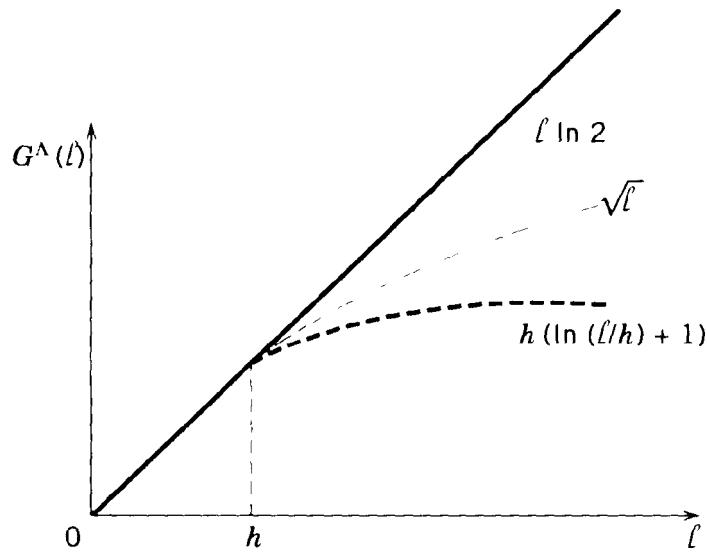


FIGURE 4.1. The growth function is either linear or bounded by a logarithmic function. It cannot, for example, behave like the dashed line.

or (b) is bounded by the inequality

$$G^A(\ell) \begin{cases} = \ell \ln 2 & \text{if } \ell \leq h \\ \leq \ln \left(\sum_{i=0}^h C_\ell^i \right) \leq \ln \left(\frac{e\ell}{h} \right)^h = h \left(1 + \ln \frac{\ell}{h} \right) & \text{if } \ell > h. \end{cases} \quad (4.45)$$

where h is the largest integer for which

$$G^A(h) = h \ln 2.$$

In other words the function $G^A(\ell)$ can be either linear or bounded by a logarithmic function with coefficient h . (It cannot, for example, be of the form $G(\ell) = \sqrt{\ell}$ (Fig 4.1).)

This theorem can be formulated in the following equivalent form, where instead of growth function one considers maximal subsets of a set of some elements.

Theorem 4.3a. Let Z be an (infinite) set of elements z and let S be some set of subsets A of the set Z . Denote by $N^S(z_1, \dots, z_\ell)$ the number of different subsets

$$(z_1, \dots, z_\ell) \cap A, \quad A \in S,$$

of the set z_1, \dots, z_ℓ . Then either

$$\sup_{z_1, \dots, z_\ell} N^S(z_1, \dots, z_\ell) = 2^\ell$$

or

$$N^S(z_1, \dots, z_\ell) \begin{cases} = 2^\ell & \text{if } \ell \leq h, \\ \leq \left(\sum_{i=0}^h C_\ell^i \right)^h \leq \left(\frac{e\ell}{h} \right)^h & \text{if } \ell > h, \end{cases}$$

where h is the last integer ℓ for which the equality is valid.

We will prove this theorem in the next section.

Theorem 4.3 asserts that sets of indicator functions can be split into two different categories:

1. Sets of indicator functions with linear growth functions
2. Sets of indicator functions with logarithmic growth functions

Definition. The capacity of a set of functions with logarithmic bounded growth function can be characterized by the coefficient h . The coefficient h is called the *VC dimension of a set of indicator functions*.[†] It characterizes the capacity of a set of functions. When the growth function is linear the VC dimension is defined to be infinite.

Below we give an equivalent definition of the VC dimension of a set of indicator functions that stress the constructive method of estimating the VC dimension.

Definition. The VC dimension of a set of indicator functions $Q(z, a)$, $a \in A$, is equal to the largest number h of vectors z_1, \dots, z_ℓ that can be separated into two different classes in all the 2^h possible ways using this set of functions (i.e., the VC dimension is the maximum number of vectors that can be *shattered* by the set of functions).

If for any n there exists a set of n vectors that can be shattered by the functions $Q(z, a)$, $a \in A$, then the VC dimension is equal to infinity.

Therefore to estimate VC dimension of the set of functions $Q(z, a)$, $a \in A$, it is sufficient to point out the maximal number h of vectors z_1^*, \dots, z_ℓ^* that can be shattered by this set of functions.

According to Theorem 4.3 if a set of functions $Q(z, a)$, $a \in A$, has finite VC dimension the growth function can be bounded using inequality (4.45).

In Section 4.11 we shall calculate the VC dimension for some sets of functions. In the remaining part of this section we show that VC dimension plays a fundamental part in obtaining a *constructive* distribution-free bound for evaluating the risk functional from empirical data (bounds which do not depend on the unknown probability measure $F(z)$) and in solving the generalized Glivenko–Cantelli problem.

[†] Abbreviation for the Vapnik–Chervonenkis dimension.

4.9.2 Constructive Distribution-Free Sounds on Generalization Ability

First we obtain constructive distribution-free conditions for uniform convergence.

Theorem 4.4. *For a set of indicator functions $Q(z, \mathbf{a})$, $\mathbf{a} \in \Lambda$, with finite VC dimension h the following two inequalities hold true:*

1. *The inequality estimating the rate of two-sided uniform convergence*

$$\begin{aligned} P \left\{ \sup_{\alpha \in \Lambda} \left| \int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \right| > \varepsilon \right\} \\ < 4 \exp \left\{ \left(\frac{h(1 + \ln(2\ell/h))}{\ell} - \varepsilon_*^2 \right) \ell \right\}, \end{aligned} \quad (4.46)$$

where $\varepsilon^* = (\varepsilon - 1/\ell)$, and

2. *The inequality estimating the rate of relative uniform convergence:*

$$\begin{aligned} P \left\{ \sup_{\alpha \in \Lambda} \frac{\left| \int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \right|}{\sqrt{\int Q(z, \alpha) dF(z)}} > \varepsilon \right\} \\ < 4 \exp \left\{ \left(\frac{h(1 + \ln(2\ell/h))}{\ell} - \frac{\varepsilon^2}{4} \right) \ell \right\}. \end{aligned} \quad (4.47)$$

To prove this theorem it is sufficient to note that

$$H_{\text{ann}}^{\Lambda, P}(\ell) \leq G^{\Lambda}(\ell) < h \left(1 + \ln \frac{\ell}{h} \right)$$

and then to use this inequality in the bounds obtained in Theorem 4.1 and Theorem 4.2.

The bounds (4.46). (4.47) provide constructive distribution-free bounds on the generalization ability of a learning machine that minimizes the empirical risk functional.

With probability $1 - \eta$ the risk for the function $Q(z, \alpha_\ell)$ which minimizes the empirical risk functional satisfies the inequality

$$R(\alpha_\ell) < R_{\text{emp}}(\alpha_\ell) + \frac{\mathcal{E}(\ell)}{2} \left(1 + \sqrt{1 + \frac{4R_{\text{emp}}(\alpha_\ell)}{\mathcal{E}(\ell)}} \right), \quad (4.48)$$

where

$$\mathcal{E}(\ell) = 4 \frac{h(\ln 2\ell/h + 1) - \ln \eta/4}{\ell}$$

With probability $1 - 2\eta$ the difference between the attained risk and the minimal one satisfies the inequality

$$\Delta(\alpha_\ell) < \sqrt{\frac{-\ln \eta}{2\ell}} + \frac{\mathcal{E}(\ell)}{2} \left(1 + \sqrt{1 + \frac{4R_{\text{emp}}(\alpha_\ell)}{\mathcal{E}(\ell)}} \right).$$

4.9.3 Solution of Generalized Glivenko-Cantelli Problem

The result obtained in Theorem 4.4 (inequality (4.46)) can be also formulated in the terms of the Generalized Glivenko–Cantelli problem: The finiteness of the VC dimension of a set of functions $Q(z, \alpha), \alpha \in A$ (set of events $\mathbf{A}_+ = \{z : Q(z, \alpha) = 1\}$), is sufficient for existence of distribution-free exponential bounds on the rate of uniform convergence.

The next theorem reinforces this result: It shows that finiteness of the VC dimension provides not only sufficient conditions for uniform convergence, but necessary conditions as well. Therefore finiteness of VC dimension of a set of functions gives the necessary and sufficient conditions for solution of the Generalized Glivenko–Cantelli problem.

Theorem 4.5. *For existence of uniform convergence of frequencies to their probabilities over a set of events $\mathbf{A}_+ = \{z : Q(z, \alpha) = 1\}$, $\alpha \in A$, with respect to any probability measure $F(z)$ it is necessary and sufficient that the set of functions $Q(z, \alpha), \alpha \in A$, has a finite VC dimension.*

If VC dimension of the set of functions $Q(z, \alpha), \alpha \in A$, is finite, then the inequality (4.46) holds true.

Proof The proof of sufficiency of the conditions of this theorem follows from Theorem 4.4.

To prove the necessity of this condition, we show that any single set of points in a space $Z(\ell)$ is measurable and a given set of functions has infinite VC dimension—that is, if for any ℓ the equality

$$\sup_{z_1, \dots, z_\ell} N^\Lambda(z_1, \dots, z_\ell) = 2^\ell \quad (4.49)$$

holds true—then for any ℓ and for any ε a probability measure $F(z)$ can be chosen such that with probability one the inequality

$$\sup_{\alpha \in \Lambda} \left| \int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \right| > 1 - \varepsilon$$

is valid.

Indeed, let us choose an integer $n > \ell/\varepsilon$. Since for any ℓ the equality (4.49) is valid, it is possible to choose n points

$$Z^n = z_1, \dots, z_n,$$

which can be shattered by functions in the set $Q(z, \alpha)$, $\alpha \in A$.

Now let us specify the probability measure: The distribution is concentrated on these points, and all points have equal probability $P(z_i) = 1/n$.

Let us consider the random sample $Z^\ell = z_1, \dots, z_\ell$ of size ℓ . Denote by Z^* the subset of Z^n that contains the points of the set Z^n not included in the set Z^ℓ . It is clear that the number of these points is not less than $n - \ell$. Since

$$N^A(z_1, \dots, z_n) = 2^n$$

there exists a function $Q(z, \alpha^*)$ that takes the value one on the vectors from the subset Z^* and the value zero on the vectors from the subset Z^ℓ . This means that

$$\frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha^*) = 0$$

and at the same time

$$\int Q(z, \alpha^*) dF(z) \geq \frac{n - \ell}{n} > 1 - \varepsilon$$

Therefore with probability one

$$\left| \int Q(z, \alpha^*) dF(z) - \sum_{i=1}^{\ell} Q(z_i, \alpha^*) \right| > 1 - \varepsilon.$$

The theorem is thus proved.

4.10 PROOF OF THEOREM 4.3

The proof of Theorem 4.3 is based on the following three lemmas.

Lemma 4.3. *If for some sequence z_1, \dots, z_ℓ and some n*

$$N^A(z_1, \dots, z_\ell) > \sum_{i=0}^{n-1} C_\ell^i,$$

then there exists a subsequence z_1^, \dots, z_n^* (of this sequence) of length n such that*

$$N^A(z_1^*, \dots, z_n^*) = 2^n.$$

Proof: Denote

$$\sum_{i=0}^{n-1} C_\ell^i = \Phi(n, \ell)$$

(here and later we denote $C_\ell^i = 0$ for $i > n$). For this function, as it is easy to verify, the relations

$$\begin{aligned}\Phi(1, \ell) &= 1, \\ \Phi(n, \ell) &= 2^\ell, \quad \text{if } \ell \leq n+1, \\ \Phi(n, \ell) &= \Phi(n, \ell-1) + \Phi(n-1, \ell-1), \quad \text{if } n \geq 2\end{aligned}\tag{4.50}$$

are valid. These relations uniquely determine the function $\Phi(n, \ell)$ for $n > 0$ and $\ell > 0$.

We shall prove the lemma by induction on n and ℓ .

- For $n = 1$ and any $\ell \geq 1$ the assertion of the lemma is obvious. Indeed, for this case

$$N^\Lambda(z_1, \dots, z_\ell) > 1$$

implies that an element z^* of the sequence exists such that for some function $Q(z, \alpha_1)$ we have

$$Q(z^*, \alpha_1) = 1,$$

while for some other function $Q(z, \alpha_2)$ we have

$$Q(z^*, \alpha_2) = 0.$$

Consequently,

$$N^\Lambda(z^*) = 2.$$

- For $n < 1$ assertion of this lemma is valid because the premise is false. Indeed, in this case the premise is

$$N^\Lambda(z_1, \dots, z_\ell) > 2^\ell,$$

which is impossible because

$$N^\Lambda(z_1, \dots, z_\ell) \leq 2^\ell.$$

- Finally, assume that the lemma is valid for $n \leq n_0$ for all n . Consider the case $n = n_0 + 1$. We show that the lemma is valid in this case also for all ℓ .

We fix $n = n_0 + 1$ and carry out the induction on ℓ . As was pointed out, for $\ell < n_0 + 1$ the lemma is valid. We assume that it is valid for $\ell \leq \ell_0$ and

show that it is valid for $\ell = \ell_0 + 1$. Indeed, let the condition of the lemma

$$N^\Lambda(z_1, \dots, z_{\ell_0}, z_{\ell_0+1}) > \Phi(n_0 + 1, \ell_0 + 1)$$

be satisfied for some sequence $z_1, \dots, z_{\ell_0}, z_{\ell_0+1}$. The lemma will be proved if we will find a subsequence of length $n_0 + 1$, say z_1, \dots, z_{n_0+1} , such that

$$N^\Lambda(z_1, \dots, z_{n_0+1}) = 2^{n_0+1}.$$

Consider subsequence z_1, \dots, z_{ℓ_0} . Two cases are possible:

Case 1:

$$N^\Lambda(z_1, \dots, z_{\ell_0}) > \Phi(n_0 + 1, \ell_0).$$

Case 2:

$$N^\Lambda(z_1, \dots, z_{\ell_0}) \leq \Phi(n_0 + 1, \ell_0).$$

In case 1, in view of the induction assumption, there exists a subsequence of length $n_0 + 1$ such that

$$N^\Lambda(z_1, \dots, z_{n_0+1}) = 2^{n_0+1}.$$

This proves the lemma in the case 1.

In case 2 we distinguish two types of subsequences of the sequence z_1, \dots, z_{ℓ_0} . We assign subsequences z_{i_1}, \dots, z_{i_r} to the first type if in the set of functions $Q(z, a), a \in A$, there exists both a function $Q(z, a^*)$ satisfying the conditions

$$Q(z_{\ell_0+1}, a^*) = 1,$$

$$\begin{aligned} Q(z_{i_k}, a^*) &= 1, & k &= 1, 2, \dots, r, \\ Q(z_j, a^*) &= 0, & \text{if } z_j &\notin \{z_{i_1}, \dots, z_{i_r}\}, \end{aligned}$$

and a function $Q(z, a^{**})$ satisfying the conditions

$$Q(z_{\ell_0+1}, a^{**}) = 0,$$

$$\begin{aligned} Q(z_{i_k}, a^{**}) &= 1, & k &= 1, 2, \dots, r, \\ Q(z_j, a^{**}) &= 0, & \text{if } z_j &\notin \{z_{i_1}, \dots, z_{i_r}\}. \end{aligned}$$

We assign subsequence z_{i_1}, \dots, z_{i_r} to the second type, if either in the set of functions $Q(z, a), a \in A$, there exists a function $Q(z, a^*)$ satisfying the conditions

$$Q(z_{\ell_0+1}, a^*) = 1,$$

$$\begin{aligned} Q(z_{i_k}, a^*) &= 1, & k &= 1, 2, \dots, r \\ Q(z_j, a^*) &= 0 & \text{if } z_j &\notin \{z_{i_1}, \dots, z_{i_r}\}, \end{aligned}$$

or there exists a function $Q(z, a^{**})$ satisfying the conditions

$$Q(z_{\ell_0+1}, a^{**}) = 0,$$

$$\begin{aligned} Q(z_{i_k}, a^{**}) &= 1, & k &= 1, 2, \dots, r, \\ Q(z_j, a^{**}) &= 0, & \text{if } z_j &\notin \{z_{i_1}, \dots, z_{i_r}\} \end{aligned}$$

(but not both).

Denote the number of subsequences of the first type by K_1 and number of subsequences of the second type by K_2 . It is easy to see that

$$N^A(z_1, \dots, z_{\ell_0}) = K_1 + K_2,$$

$$N^A(z_1, \dots, z_{\ell_0}, z_{\ell_0+1}) = 2K_1 + K_2,$$

and hence

$$N^A(z_1, \dots, z_{\ell_0}, z_{\ell_0+1}) = N^A(z_1, \dots, z_{\ell_0}) + K_1. \quad (4.51)$$

Denote by $Q(z, a)$, $a \in A^*$, the subset of set of functions $Q(z, a)$, $a \in A$, that on $z_1, \dots, z_{\ell+1}$ induces the subsequences of the first type. If

$$K_1 = N^{A^*}(z_1, \dots, z_{\ell_0}) > \Phi(n_0, \ell_0),$$

then, in view of induction hypothesis, there exists a subsequence $z_{i_1}, \dots, z_{i_{n_0}}$ such that

$$N^{A'}(z_{i_1}, \dots, z_{i_{n_0}}) = 2^{n_0}.$$

However, in this case

$$N^{A^*}(z_{i_1}, \dots, z_{i_{n_0}}, z_{\ell_0+1}) = 2^{n_0+1}$$

for sequence $z_{i_1}, \dots, z_{i_{n_0}}, z_{\ell_0+1}$, since this subsequence belongs to the subsequence of the first type.

If, however,

$$K_1 = N^{A^*}(z_1, \dots, z_{\ell_0}) \leq \Phi(n_0, \ell_0) \quad (4.52)$$

we obtain in view of (4.51) and (4.52)

$$N^A(z_1, \dots, z_{\ell_0+1}) \leq \Phi(n_0 + 1, \ell_0) + \Phi(n_0, \ell_0),$$

which, by virtue of the properties (4.50) of the function $\Phi(n, \ell)$, implies that

$$N^A(z_1, \dots, z_{\ell_0+1}) \leq \Phi(n_0 + 1, \ell_0 + 1).$$

This contradicts the condition of the lemma.

The lemma is proved.

Lemma 4.4. *If for some n*

$$\sup_{z_1, \dots, z_{n+1}} N^A(z_1, \dots, z_{n+1}) \neq 2^{n+1},$$

then *for all $\ell > n$ the inequality*

$$\sup_{z_1, \dots, z_\ell} N^A(z_1, \dots, z_\ell) \leq \Phi(n + 1, \ell)$$

holds true.

Proof Let $\sup_{z_1, \dots, z_\ell} N^\Lambda(z_1, \dots, z_\ell)$ not be identically equal to 2^ℓ , and let $n+1$ be the first value of ℓ such that

$$\sup_{z_1, \dots, z_{n+1}} N^\Lambda(z_1, \dots, z_{n+1}) \neq 2^{n+1}.$$

Then for any sample of size ℓ , larger than n , the equality

$$N^\Lambda(z_1, \dots, z_\ell) \leq \Phi(n+1, \ell)$$

is valid. Indeed, otherwise, in view of Lemma 4.3, one could find the subsequence $z_{i_1}, \dots, z_{i_{n+1}}$ such that

$$N^\Lambda(z_{i_1}, \dots, z_{i_{n+1}}) = 2^{n+1},$$

which is impossible because by assumption, $\sup_{z_1, \dots, z_{n+1}} N^\Lambda(z_1, \dots, z_{n+1}) \neq 2^{n+1}$.

The lemma is proved.

Lemma 4.5. *For $\ell > n$ the following bound is true:*

$$\Phi(n, \ell) < 1.5 \frac{\ell^{n-1}}{(n-1)!} < \left(\frac{e\ell}{n-1} \right)^{n-1}. \quad (4.53)$$

Proof Since the relation (4.50) is fulfilled for $\Phi(n, \ell)$, to prove (4.53) it is sufficient to verify that for $\ell > n$ the inequality

$$\frac{e^{n-1}}{(n-1)} + \frac{\ell^n}{n!} \leq \frac{(\ell+1)^n}{n!} \quad (4.54)$$

is valid and to verify (4.54) on the boundary (i.e., $n=1$, $\ell=n+1$).

The inequality (4.54) is clearly equivalent to inequality

$$\ell^{n-1}(\ell+n) - (\ell+1)^n \leq 0,$$

whose validity follows from Newton's binomial expansion.

It thus remains to verify (4.54) on the boundary. For $n=1$ the verification is direct. Next we shall verify the bound for small values of n and ℓ :

$\ell = n+1$	2	3	4	5	6
$\Phi(n, \ell)$	1	4	11	26	57
$1.5 \frac{\ell^{n-1}}{(n-1)!}$	1.5	4.5	12	31.25	81

To check (4.54) for $n > 6$ we utilize Stirling's formula for an upper bound on $\ell!$

$$\ell! < \sqrt{2\pi\ell}\ell^\ell \exp\{\ell - (12\ell)^{-1}\},$$

where for $\ell = n + 1$ we obtain

$$\frac{\ell^{n+1}}{(n-1)!} = \frac{(\ell-1)\ell^{\ell-1}}{\ell!} \geq \frac{\ell-1}{\sqrt{2\pi\ell}\ell} \exp\left\{-\ell + \frac{1}{12\ell}\right\}$$

and for $\ell > 6$ we have

$$\frac{\ell^{(n+1)}}{(n-1)!} \geq 0.8 \frac{1}{\sqrt{2\pi\ell}} e^{-\ell}.$$

On the other hand, $\Phi(n, \ell) \leq 2^\ell$ always. Therefore it is sufficient to verify that for $\ell \geq 6$

$$2^\ell \leq 1.2\sqrt{2\pi\ell}e^\ell.$$

Actually it is sufficient to verify the inequality for $\ell = 6$ (which is carried out directly) since as ℓ increases the right-hand side of inequality grows faster than the left-hand side.

The lemma is proved.

The assertions of Lemmas 4.4 and 4.5 imply the validity of Theorem 4.3.

4.11 EXAMPLE OF THE VC DIMENSION OF THE DIFFERENT SETS OF FUNCTIONS

In this section we give several examples of estimating the VC dimension of different sets of functions.

According to Theorem 4.3 if the VC dimension of a set of indicator functions is finite, the inequality

$$\max_{z_1, \dots, z_\ell} N^\Lambda(z_1, \dots, z_\ell) \leq \sum_{i=0}^h C_\ell^i \quad (4.55)$$

holds true, where h is the maximal number $\ell = h$ such that

$$\max_{z_1, \dots, z_h} N^\Lambda(z_1, \dots, z_h) = 2^h.$$

First of all we give a simple example for which (4.55) turns out an equality. This implies that the general result (4.55) cannot be improved.

Example 1 (The obtained bound of the growth function is tight). Let Z be an arbitrary set and let S be a set of subsets of Z , such that every $A \in S$ contains less than h elements. Consider a set of indicator functions $Q(z, a)$, $a \in A$, determined on S such that for any subset A of Z the function $Q(z, \alpha(A))$ is one on the elements of A and is zero on the elements $Z - A$. For this set of functions

$$\max_{z_1, \dots, z_\ell} N^A(z_1, \dots, z_\ell) = 2^h \quad \text{if } \ell \leq h$$

and

$$\max_{z_1, \dots, z_\ell} N^A(z_1, \dots, z_\ell) = \sum_{i=0}^h C_\ell^i \quad \text{if } \ell > h.$$

Example 2 (The VC dimension of a set of functions linear in their parameters is equal to the number of parameters). Consider a set of indicator functions linear in their parameters:

$$Q(z, \alpha) = \theta \left(\sum_{k=1}^n a^k \phi_k(z) \right), \quad \alpha = (a^1, \dots, a^n), \quad a^i \in (-\infty, \infty). \quad (4.56)$$

We shall show that the VC dimension of this set of functions equals n , the number of free parameters (we assume that $\phi_k(z)$, $k = 1, \dots, n$, is a set of linearly independent functions).

To prove this we denote $u^k = \phi_k(z)$, $k = 1, 2, \dots, n$, and consider the set of linear indicator functions $l(u, a)$ passing through the origin in the space $U = (u^1, \dots, u^n)$

$$l(u, \alpha) = \theta \left(\sum_{k=1}^n a^k u^k \right). \quad (4.57)$$

It is clear that the maximal number of different separations of ℓ vectors from Z using the set of functions (4.56) is equal to the maximal number of different separations of ℓ vectors from U using the set of functions (4.57).

Thus let us estimate a number of different separations of ℓ vectors from U using the hyperplanes passing through the origin in the n -dimensional space U . It is easy to see that the following n vectors from R^n

$$(1, 0, \dots, 0), (0, 1, 0, \dots, 0), \dots, (0, \dots, 0, 1)$$

can be shattered by hyperplanes

$$(u * a) = 0,$$

(here we denote by $(u * a)$ the inner product of two vectors in R^n).

Let us show that there are no $n+1$ vectors in R^n that can be shattered by hyperplanes passing through the origin. Suppose the contrary: Let u_1, \dots, u_{n+1}

be vectors that can be shattered. This implies that there exist 2^{n+1} vectors $a_i \in R^n$, $i = 1, \dots, 2^{n+1}$ which form a $(n \times 1) \times 2^{n+1}$ matrix of inner products $z_{i,j} = (u_i * a_j)$, $i = 1, \dots, (n+1)$, $j = 1, \dots, 2^{n+1}$

$$A = \begin{vmatrix} z_{1,1} & \cdots & z_{1,2^{n+1}} \\ \vdots & \cdots & \vdots \\ \vdots & \cdots & \vdots \\ z_{(n+1),1} & \cdots & z_{(n+1),2^{n+1}} \end{vmatrix}$$

The elements $z_{i,j}$ of this matrix are such that 2^{n+1} columns of the matrix have all 2^{n+1} possible combination of signs

$$\text{sign}(A) = \begin{vmatrix} - & - & - & - & + \\ \vdots & \cdot & \cdot & \cdot & + \\ \vdots & \cdot & \cdot & \cdot & + \\ - & + & \cdot & \cdot & + \end{vmatrix}$$

Therefore, row-vectors $Z_i = (z_{i,1}, \dots, z_{i,2^{n+1}}, i = 1, \dots, (n+1))$, of A-matrix are linearly independent since there are no constants c_1, \dots, c_{n+1} such that

$$\sum_{i=1}^{n+1} c_i Z_i = 0$$

because for any constants c_1, \dots, c_{n+1} there is a column with the same signs. This implies that $n+1$ vectors u_1, \dots, u_{n+1} in R^n are linearly independent and this contradiction proves that there are no $n+1$ vectors in R^n that can be shattered by hyperplanes passing through the origin. Therefore the maximum number of vectors that can be shattered by hyperplanes passing through the origins is n and consequently the VC dimension of this set of functions is n .

Now we show that the bound on the growth function for a set of linear hyperplanes that follows from Theorem 4.3 is rather accurate. To show this, let us estimate the value of $\max_{u_1, \dots, u_\ell} N^A(u_1, \dots, u_\ell)$.

To do this, note that to any vector $u = (u^1, \dots, u^n)$ of the space U there corresponds a hyperplane

$$\sum_{i=1}^n a^k u_*^k = 0$$

in the space $A = (a^1, \dots, a^n)$. And vice versa to any vector $a = (a_*^1, \dots, a_*^n)$ of the space A corresponds hyperplane

$$\sum_{i=1}^n a_*^k u^k = 0$$

in the space U .

Thus to ℓ vectors $u_i, i = 1, \dots, \ell$ in the space U there correspond ℓ hyperplanes passing through the origin in the space A.

Our assertion is the following: The maximal number of different separations of ℓ vectors by hyperplanes passing through origin in the space U is equal to the number of different components into which the ℓ hyperplanes separate the n-dimensional space A (see Fig. 4.2).

Indeed, let I - be a vector in A corresponding to some hyperplane in U . If one continuously rotates this hyperplane in the space U such that separation of z_1, \dots, z_ℓ remains in fact, the corresponding trajectory of the vector Γ belongs to the same component of the space A.

We shall estimate the number of different components into which ℓ hyperplanes can divide the n-dimensional space. Let us denote by $\Phi(n, \ell)$ the maximal number of components into which ℓ hyperplanes can divide the n-dimensional space. Let us determine a recurrent procedure for estimating the number of components.

It is clear that in the one-dimensional case for a hyperplane passing through the origin we have

$$\Phi(1, \ell) = 2.$$

One hyperplane divides any space into two components

$$\Phi(n, 1) = 2.$$

Now, let $\ell - 1$ hyperplanes $\Gamma_1, \dots, \Gamma_{\ell-1}$ divide n-dimensional space into $\Phi(n, \ell - 1)$ components. Let us add one new hyperplane Γ_ℓ .

If this hyperplane passes through one of the "old" components, then it divides this component into two parts. Otherwise, the old component is preserved.

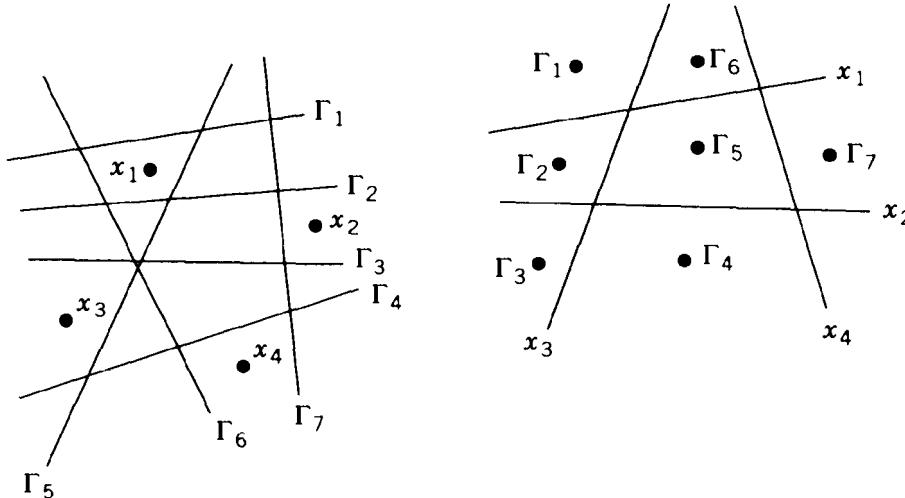


FIGURE 4.2. To any vector $u_j, j = 1, \dots, \ell$ in the space U there correspond hyperplanes passing through the origin in the space A.

Thus, if one added a new hyperplane Γ_ℓ , the number of components can be increased by the quantity equal to the number of components which are split by this hyperplane. Conversely, any component K_i makes a trace $K_i \cap \Gamma_\ell$ on Γ_ℓ . The number of these traces is equal to the number of parts in which $n - 1$ hyperplanes $\Gamma_1, \dots, \Gamma_{\ell-1}$ divide the hyperplane Γ_ℓ .

Since the dimensionality of Γ_ℓ is equal to $n - 1$ the number of traces does not exceed $\Phi(n - 1, \ell - 1)$. Thus we obtain the following recurrent equation:

$$\Phi(n, \ell) = \Phi(n, \ell - 1) + \Phi(n - 1, \ell - 1), \quad (4.58)$$

$$\begin{aligned} \Phi(n, 1) &= 2, \\ \Phi(1, \ell) &= 2. \end{aligned}$$

The solution of Eq. (4.58) is

$$\Phi(n, \ell) = \begin{cases} 2^\ell & \text{if } n > \ell, \\ 2 \sum_{i=0}^{n-1} C_{\ell-1}^i & \text{if } n \leq \ell. \end{cases} \quad (4.59)$$

Note that according to the exact formula (4.59) the growth function for a set of linear functions in the region $\ell > n$ is equal to

$$G^\Lambda(\ell) = \ln \left(2 \sum_{i=0}^{n-1} C_{\ell-1}^i \right) < (n - 1) \left(\ln \frac{\ell}{n-1} + 1 \right) + \ln 2.$$

The bound for the growth function for region $\ell > n$ obtained in Theorem 4.3 is equal to

$$G^\Lambda(\ell) \leq \ln \left(\sum_{i=0}^n C_\ell^i \right) < n \left(\ln \frac{\ell}{n} + 1 \right)$$

One can see how close the bound is to the exact result in this case.

The next two examples show that VC dimension of a set of indicator functions that have nonlinear dependence on parameters can differ from the number of parameters.

Example 3: The VC dimension of the set of indicator functions nonlinear in parameters can be less than the number of parameters. Let us consider the following set of one-dimensional functions

$$Q(z, \alpha) = \theta \left(\sum_{d=1}^n |a_d z^d| \operatorname{sign} z + a_0 \right), \quad a_d \in R^1.$$

This set of functions is a set of monotonic nondecreasing indicator functions. It is clear that using a set of monotonic nondecreasing indicator functions on

the line one can shatter only one point. This means that VC dimension of the set of functions considered here is independent of the number of parameters n .

Example 4: The VC dimension of the set nonlinear in parameters indicator functions can exceed the number of parameters. Lastly consider the following set of one-dimensional indicator functions

$$Q(z, \alpha) = \theta(\sin \alpha z), \quad z \in (0, 2\pi), \quad \alpha \in (0, \infty)$$

defined on the interval $(0, 2\pi)$.

We show that the VC dimension of this set of functions equals to infinity if we establish that for any ℓ and any binary sequence

$$\delta_1, \dots, \delta_\ell, \quad \delta_i \in \{0, 1\}$$

there exist ℓ points z_1, \dots, z_ℓ such that the system of equation

$$\theta(\sin \alpha z_i) = \delta_i, \quad i = 1, 2, \dots, \ell \quad (4.60)$$

has a solution in α . Let us consider the points $z_i = 2\pi 10^{-i}$, $i = 1, 2, \dots, \ell$. It is easy to check that for these points the value

$$\alpha^* = \frac{1}{2} \left(\sum_{i=1}^{\ell} (1 - \delta_i) 10^i + 1 \right)$$

gives a solution of the system of equations (4.60).

Thus in general the number of parameters does not determine the VC dimension of a set of functions. But it is the VC dimension rather than the number of parameters of the set of functions that defines the generalization ability of a learning machine. This fact will play an extremely important role in constructing learning algorithms later. Chapter 10 introduces learning machines that realize functions with low VC dimension and have billions of parameters.

4.12 REMARKS ABOUT THE BOUNDS ON THE GENERALIZATION ABILITY OF LEARNING MACHINES

Thus in this chapter we obtained the bounds on the generalization ability of learning machines that minimize the empirical risk functional. These bounds can be described as follows:

With probability at least $1 - \eta$ the inequality

$$R(\alpha_\ell) \leq R_{\text{emp}}(\alpha_\ell) + \frac{\mathcal{E}(\ell)}{2} \left(1 + \sqrt{1 + \frac{R_{\text{emp}}(\alpha_\ell)}{\mathcal{E}(\ell)}} \right) \quad (4.61)$$

holds true.

With probability at least $1 - 2\eta$ the inequality

$$\Delta(\alpha_\ell) = \frac{\mathcal{E}(\ell)}{2} \left(1 + \sqrt{1 + \frac{R_{\text{emp}}(\alpha_\ell)\ell}{\mathcal{E}(\ell)}} \right) + \sqrt{\frac{-\ln \eta}{\ell}} \quad (4.62)$$

holds true.

Different expressions for $\mathcal{E}(\ell)$ define different types of bounds. The expression

$$\mathcal{E}(\ell) = 4 \frac{H_{\text{ann}}^\Lambda(2\ell) - \ln \eta/4}{\ell},$$

where $H_{\text{ann}}^\Lambda(\ell)$ is the annealed entropy, defines tight distribution dependent bounds that are valid for a specific learning machine (a specific set of functions) and a specific problem (a specific probability measure).

One can exclude information about the probability measure by using the expression

$$\mathcal{E}(\ell) = 4 \frac{G^\Lambda(2\ell) - \ln \eta/4}{\ell},$$

where $G^\Lambda(\ell)$ is the growth function of a set of functions $Q(z, a), a \in \mathbf{A}$. Bounds (4.61), (4.62) with this expression for $\mathcal{E}(\ell)$ are valid for a given learning machine and any problem (any probability measure).

These bounds are to be thought conceptual rather than constructive since the theory does not give a regular way for estimating the annealed entropy or growth function. Therefore we use the upper bound of the growth function that is based on the VC dimension of the set of functions. (Theorem 4.3 points out a constructive way for evaluating the VC dimension of a set of functions.) The constructive bounds are based on the following expression for $\mathcal{E}(\ell)$:

$$\mathcal{E}(\ell) = 4 \frac{h(\ln(2\ell/h) + 1) - \ln \eta/4}{e}$$

Let us denote

$$\tau = \frac{\ell}{h};$$

then the bound for the $\mathcal{E}(\ell)$ is

$$\mathcal{E}(\ell) \leq 4 \frac{\ln 2\tau + 1}{\tau} - \frac{\ln \eta/4}{\ell}.$$

This expression shows that the generalization ability of a learning machine depends on the ratio of the number of observations to the VC dimension of the set of functions (for reasonable η the second term in the expression is negligibly small compared to the first one).

An important goal of the theory is to find a more accurate constructive bound than the one described. According to the Key theorem proved in Chapter 3, the uniform convergence forms the necessary and sufficient conditions for consistency of the ERM method. Therefore to obtain more accurate bounds on the rate of the learning processes based on the ERM method, one has to obtain a more accurate bound on the rate of uniform convergence.

To construct any bound, one has to use some capacity concept. From the conceptual point of view the accuracy of the obtained bound depends on which type of capacity concept is used. We obtained the best bound using the annealed entropy concept. However, the construction of this concept uses the unknown distribution function $F(z)$.

The bounds obtained on the basis of the growth function concept or the VC dimension concept are another extreme case: They ignore any a priori information about unknown distribution function $F(z)$.

It is very important to find the way how to obtain ***constructive*** bounds using general information about the unknown distribution function $F(z)$. The nonconstructive bound can be obtained easily.

Indeed, suppose one has the information about the unknown probability measure $F(z) \in \mathcal{P}$, where \mathcal{P} is some set of densities. Then one can immediately suggest tight nonconstructive distribution-dependent bounds based on the following generalized growth function concept:

$$M_{\mathcal{P}}^{\Lambda}(\ell) = \sup_{F(z) \in \mathcal{P}} E \ln N^{\Lambda}(z_1, \dots, z_{\ell}).$$

Since

$$H_{\text{ann}}^{\Lambda}(\ell) \leq M_{\mathcal{P}}^{\Lambda}(\ell) \leq G^{\Lambda}(\ell)$$

the bounds (4.61), (4.62) with

$$\mathcal{E}(\ell) = 4 \frac{M_{\mathcal{P}}^{\Lambda}(2\ell) - \ln \eta/4}{\ell}$$

are valid. These bounds are not based on the knowledge of the specific distribution function $F(z)$; however, they take into account a priori information about the set that includes this function. Therefore these nonconstructive bounds are tighter than nonconstructive bounds based on the growth function.

To develop constructive bounds, one has to find a constructive bound for the generalized growth function that is better than the one based on the VC dimension. The main problem here is to find some set of probability measures \mathcal{P} for which one can obtain a constructive bound on the Generalized Growth

function just as constructive distribution-free bounds were obtained using the VC dimension.

For the theory of distribution free bounds only one question remains:

How tight are the obtained bounds?

The Appendix to this chapter tries to answer this question. We give lower bounds for the generalization ability of algorithms which minimize empirical risk. These lower bounds are reasonably close to the upper bounds derived in this chapter. This will ensure that the theory of bounds constructed in this chapter is rather tight.

4.13 BOUND ON DEVIATION OF FREQUENCIES IN TWO HALF-SAMPLES

In proving the basic inequalities (Theorem 4.1, and Theorem 4.2) we use the bound for the deviation of frequencies in two half-samples. In this section we show how this bound is obtained.

Our goal is to estimate the value

$$\Gamma = \sum_k \frac{C_m^k C_{2\ell-m}^{\ell-k}}{C_{2\ell}^{\ell}},$$

where the summation is conducted over k so that

$$\left| \frac{k}{\ell} - \frac{m-k}{\ell} \right| > \varepsilon, \quad \max(0, m-\ell) \leq k \leq \min(m, \ell)$$

and where ℓ and $m < 2\ell$ are arbitrary positive integers.

The last inequality is equivalent to inequality

$$\left| k - \frac{m}{2} \right| > \frac{\varepsilon\ell}{2}, \quad \max(0, m-\ell) \leq k \leq \min(m, \ell).$$

We decompose Γ into two summands,

$$\Gamma = \Gamma_1 + \Gamma_2,$$

where we denote

$$\Gamma_1 = \sum_k \frac{C_m^k C_{2\ell-m}^{\ell-k}}{C_{2\ell}^{\ell}}, \quad \text{where } k > \frac{\varepsilon\ell}{2} + \frac{m}{2},$$

$$\Gamma_2 = \sum_k \frac{C_m^k C_{2\ell-m}^{\ell-k}}{C_{2\ell}^{\ell}}, \quad \text{where } k < \frac{\varepsilon\ell}{2} - \frac{m}{2}.$$

We introduce the following notations

$$p(k) = \frac{C_m^k C_{2\ell-m}^{\ell-k}}{C_{2\ell}^{\ell}}, \quad (4.63)$$

$$q(k) = \frac{p(k+1)}{p(k)} = \frac{(m-k)(\ell-k)}{(k+1)(\ell+k+1-m)}, \quad (4.64)$$

where

$$\max(0, m - \ell) \leq k \leq \min(m, \ell).$$

Furthermore, we denote

$$s = \min(m, \ell), \quad T = \max(0, m - \ell);$$

$$d(k) = \sum_{i=k}^s p(i).$$

Clearly, the relation

$$d(k+1) = \sum_{i=k+1}^s p(i) = \sum_{i=k}^{s-1} p(i+1) = \sum_{i=k}^{s-1} p(i)q(i) \quad (4.65)$$

is valid. Furthermore, it follows from (4.64) that for $i < j$ we have $q(i) > q(j)$; that is, $q(i)$ is monotonically decreasing. Therefore the inequality

$$d(k+1) = \sum_{i=k}^{s-1} p(i)q(i) < q(k) \sum_{i=k}^s p(i)$$

follows from (4.65). By definition of $d(k)$ we have

$$d(k+1) < q(k)d(k).$$

Applying this relation successively, we obtain the following for arbitrary k and j such that $T \leq j < k \leq s-1$:

$$d(k) < d(j) \prod_{i=j}^{k-1} q(i).$$

Since $d(j) \leq 1$ we have

$$d(k) < \prod_{i=j}^{k-1} q(i), \quad (4.66)$$

where j is an arbitrary integer smaller than k .

We denote

$$t = k - \frac{m-1}{2}.$$

Then

$$q(t) = \frac{\frac{m+1}{2} - t}{\frac{m+1}{2} + t} \cdot \frac{\left(\ell - \frac{m-1}{2}\right) - t}{\left(\ell - \frac{m-1}{2}\right) + t}.$$

Moreover, as long as $T < k < s$, the inequality

$$|t| < \min\left(\frac{m+1}{2}, \ell - \frac{m-1}{2}\right)$$

is clearly valid.

To approximate $q(k)$ we analyze the function

$$F(t) = \frac{a-t}{a+t} \cdot \frac{b-t}{b+t},$$

assuming that a and b are both positive.

For $|t| < \min(a, b)$ we obtain

$$\ln F(t) = \ln(a-t) - \ln(a+t) + \ln(b-t) - \ln(b+t).$$

Furthermore, we have

$$\ln F(0) = 0,$$

$$\frac{d}{dt}(\ln F(t)) = -\left[\frac{2a}{a^2-t^2} + \frac{2b}{b^2-t^2}\right].$$

This implies that for $|t| < \min(a, b)$ the inequality

$$\frac{d}{dt}(\ln F(t)) \leq -2\left[\frac{1}{a} + \frac{1}{b}\right]$$

is valid. Consequently, for $|t| < \min(a, b)$ and $t \geq 0$ the inequality

$$\ln F(t)) \leq -2 \left[\frac{1}{a} + \frac{1}{b} \right] t$$

is valid.

Returning to $q(t)$ we obtain for $t \geq 0$

$$\ln q(t) \leq -2 \left[\frac{2}{m+1} + \frac{2}{2\ell-m+1} \right] t = -8 \frac{\ell+1}{(m+1)(2\ell-m+1)} t.$$

We now bound

$$\ln \left(\prod_{i=j}^{k-1} q(i) \right)$$

assuming that $(m-1)/2 \leq j \leq k-1$:

$$\begin{aligned} \ln \left(\prod_{i=j}^{k-1} q(i) \right) &= \sum_{i=j}^{k-1} \ln q(i) \\ &\leq \frac{-8(\ell+1)}{(m+1)(2\ell-m+1)} \sum_{i=j}^{k-1} \left(i - \frac{m-1}{2} \right). \end{aligned}$$

Returning to (4.66), we obtain

$$\ln d(k) < \frac{-8(\ell+1)}{(m+1)(2\ell-m+1)} \sum_{i=j}^{k-1} \left(i - \frac{m-1}{2} \right),$$

where j is an arbitrary number smaller than k . Therefore for $k > (m-1)/2$ one can set $j = (m-1)/2$ for m odd and $j = m/2$ for m even, obtaining a stronger bound. Next, summing the arithmetic progression, we obtain the inequality

$$\ln d(k) < \frac{-4(\ell+1)}{(m+1)(2\ell-m+1)} \left(k - \frac{m}{2} \right)^2$$

for even m and obtain the inequality

$$\ln d(k) < \frac{-4(\ell+1)}{(m+1)(2\ell-m+1)} \left(k - \frac{m-1}{2} \right) \left(k - \frac{m-1}{2} - 1 \right)$$

for odd m .

Finally Γ_1 is $d(k)$ for the first integer k such that

$$k - \frac{m}{2} > \frac{\varepsilon^2 \ell}{2},$$

from which we obtain

$$\ln \Gamma_1 < -\frac{\ell + 1}{(m + 1)(2\ell - m + 1)} \varepsilon^2 \ell^2. \quad (4.67)$$

The right-hand side of (4.67) attains its maximum at $m = \ell$, and consequently

$$\Gamma_1 < \exp \left\{ -\frac{\varepsilon^2 \ell^2}{\ell + 1} \right\} \approx \exp\{-\varepsilon^2 \ell\}. \quad (4.68)$$

In the same manner one can bound Γ_2 , since the distribution (4.63) is symmetric with respect to the point $k = m/2$. Since $\Gamma_1 = \Gamma_2$ we obtain

$$\Gamma < 2 \exp \left\{ -\frac{\varepsilon^2 \ell^2}{\ell + 1} \right\} \approx 2 \exp\{-\varepsilon^2 \ell\}. \quad (4.69)$$

APPENDIX TO CHAPTER 4: LOWER BOUNDS ON THE RISK OF THE ERM PRINCIPLE

Until now we have accepted the empirical risk minimization principle without any discussion. We found the bounds describing the generalization ability of this principle for sets of indicator functions.

Now we would like to discuss the following two questions:

1. Is the principle of empirical risk minimization a good one? Can it be considered as optimal in some sense?
2. How tight are the bounds obtained for the class of learning machines minimizing the empirical risk?

To answer these questions we have to determine:

1. What general strategy of statistical inference reflects the method of empirical risk minimization?
2. How close are the upper bounds obtained for the ERM principle to the lower bounds?

This appendix tries to answer these questions by showing that the ERM principle reflects the philosophy of the so-called **minimax loss** strategy (not to be confused with minimax strategy). Despite the fact that the ERM method does not guarantee the minimum of the maximum possible losses, its upper bounds are relatively close to the lower bound on the minimax loss. That is, the losses for the ERM methods are close to the minimax losses.

A4.1 TWO STRATEGIES IN STATISTICAL INFERENCE

Consider the situation where we would like to choose an algorithm $A \in \mathcal{A}$ for solving a set of problems $\pi \in \Pi$.

Suppose that for any algorithm A and for any problem π we can define the value $T(\pi, A)$ which characterizes the quality of the solution of the problem π by the algorithm A (let smaller values $T(\pi, A)$ means better quality).

The question is how to choose *one algorithm* for solving sets of problems, if for any problem π there exists its own optimal algorithm A , which minimizes $T(\pi, A)$.

For this situation the theory of statistical inference suggests two strategies, namely, the *Bayesian strategy* and the *Minimax loss* strategy.

Let the smallest loss for the problem π be $T_0(\pi)$. Consider the *loss functional*[†]

$$L(\pi, A) = T(\pi, A) - T_0(\pi),$$

which evaluates the loss in solving the problem π if instead of the best possible solution for this problem one obtains the solution provided by algorithm A .

The Bayesian strategy suggests that we choose the algorithm A_B , which minimizes the expectation of loss over all problems Π . This means that one should be given an a priori distribution $P(\pi)$ on a set of problems Π , which allows one to construct the expectation

$$L_B(A) = \int L(\pi, A) dP(\pi). \quad (\text{A4.1})$$

The minimum of functional (A4.1) determines the Bayesian algorithm A_B .

The minimax loss strategy suggests a more cautious approach. According to the minimax loss strategy, choose the algorithm A_M , which minimizes the losses for the worst (for this algorithm) problem in the set Π . In other words, choose the algorithm A_M , which minimizes the functional[‡]

$$L_M(A) = \sup_{\pi \in \Pi} L(\pi, A). \quad (\text{A4.2})$$

Denote the algorithm that minimizes (A4.2) by A_M . It is easy to verify that for any distribution function $P(\pi)$ determining the Bayesian strategy the

[†] Note that in Chapter 1 we introduced the concept of *loss functions* $Q(z, \alpha)$, $\alpha \in \Lambda$, which we used to construct *risk functional* $R(\alpha) = \int Q(z, \alpha) dF(z)$. Here we consider a new concept the *loss functional*, which is used to analyze quality of various statistical strategies for solution of a set of problems [defined by various $F(z)$].

[‡] Note that the minimax loss strategy differs from the *minimax strategy*, which suggests that we choose by algorithm A minimizing the functional

$$L_M^* = \sup_{\pi \in \Pi} T(\pi, A).$$

inequality

$$\inf_{A \in \mathcal{A}} L_B(A) \leq \inf_{A \in \mathcal{A}} L_M(A) \quad (\text{A4.3})$$

holds true.[†]

Let A_B be an algorithm that minimizes (A4.1). Then

$$\begin{aligned} L_B(A_B) &= \inf_{A \in \mathcal{A}} L_B(A) = \inf_{A \in \mathcal{A}} \int L(\pi, A) dP(\pi) \\ &\leq \int L(\pi, A_M) dP(\pi) \leq \int \sup_{\pi \in \Pi} L(\pi, A_M) dP(\pi) \\ &= \sup_{\pi \in \Pi} L(\pi, A_M) = L_M(A_M). \end{aligned}$$

This Appendix shows that the empirical risk minimization principle results in algorithms A_{emp} , which are close to the optimal ones in the sense of minimax loss strategy.

To prove this we note that

$$L_M(A_M) \leq L_M(A_{\text{emp}}).$$

Below we first find upper bounds of the maximal loss for the empirical risk minimization method. Then we derive lower bounds for the minimax loss. These two bounds turn out to be reasonably close.

A4.2 MINIMAX LOSS STRATEGY FOR LEARNING PROBLEMS

This book considers the learning problem as a problem of minimizing the risk functional

$$R(\pi, \alpha) = \int Q(z, \alpha) dF_\pi(z)$$

on the basis of empirical data

$$z_1, \dots, z_\ell.$$

In this setting the specific problem π is determined by an unknown distribution function $F_\pi(z)$ which defines the risk functional.

[†] It is important to note that inequality (A4.3) does not mean that for solving specific problems the Bayesian strategy is better than the minimax strategy. The quantity on the left-hand side of inequality (A4.3) gives the best average loss, while the quantity on the right-hand side (A4.3) gives the best guaranteed loss.

Suppose that we are given an algorithm \mathbf{A} that in order to minimize risk $R(\pi, \mathbf{a})$ using the data chooses the function described by the parameter $\alpha_{\mathbf{A}}(z_1, \dots, z_\ell)$. This notation indicates that given the data algorithm \mathbf{A} selects function $Q(z, \alpha_{\mathbf{A}}(z_1, \dots, z_\ell))$.

The value of risk for the chosen function is

$$R(\pi, \alpha_{\mathbf{A}}(z_1, \dots, z_\ell)) = \int Q(z, \alpha_{\mathbf{A}}(z_1, \dots, z_\ell)) dF_\pi(z).$$

Let the expectation

$$T(\pi, \mathbf{A}) = \int R(\pi, \alpha_{\mathbf{A}}(z_1, \dots, z_\ell)) dF_\pi(z_1, \dots, z_\ell)$$

define the quality of solution of the problem π by the algorithm \mathbf{A} using data of size ℓ .

Consider the following loss functional:

$$L(\pi, \mathbf{A}) = T(\pi, \mathbf{A}) - T(\pi). \quad (\text{A4.4})$$

Our goal is:

1. First to obtain the upper bound for the functional

$$L_M(A_{\text{emp}}) = \sup_{\pi \in \Pi} L(\pi, A_{\text{emp}})$$

2. Then to obtain a lower bound on minimax losses for the set of problems Π .

$$L_M(A_M) = \inf_{\mathbf{A}} L_M(\mathbf{A}) = \inf_{\mathbf{A}} \sup_{\pi} L(\pi, \mathbf{A})$$

Since

$$L_M(A_M) \leq L_M(A_{\text{emp}})$$

if one finds that the lower bound for $L_M(A_M)$ is close to the upper bound for $L_M(A_{\text{emp}})$, then one can conclude that the ERM method provides losses that are close to the minimax loss for a given set of problems. In any case, the lower bounds are obtained for the guaranteed generalization ability of the method of empirical risk minimization.

Below we derive both an upper bound for the loss **(A4.3)** and a lower bound for the loss **(A4.2)** for two cases:

- a Optimistic case (for set of problems Π for which $T(\pi) = 0$)
- a Pessimistic case (for set of problems Π where there are π such that $T(\pi) \neq 0$)

For the optimistic case we show that if a learning machine minimizes cm-

pirical risk in a set of functions with VC dimension h , then for $\ell > h$ the following inequalities are valid:

$$\frac{1}{2e} \frac{h+1}{\ell+1} \leq L_M(A_M) \leq L_M(A_{\text{emp}}) \leq 4 \frac{h \left(\ln \frac{2\ell}{h} + 1 \right) + 4}{\ell} + \frac{16}{\ell}. \quad (\text{A4.5})$$

For the pessimistic case we show that if a learning machine minimizes the empirical risk in a set of functions with VC dimension h , then for $\ell > 2h$ the following inequalities are valid

$$\sqrt{\frac{h}{\ell}} (1 - \text{erf}(1)) \leq L_M(A_M) \leq L_M(A_{\text{emp}}) \leq 4 \left(\sqrt{\frac{h \left(\ln \frac{2\ell}{h} + 1 \right) + 24}{\ell}} \right). \quad (\text{A4.6})$$

In the next section we derive the upper bounds in inequalities (A4.5) and (A4.6), and then in Sections A4.4 and A4.5 we derive the lower bounds.

A4.3 UPPER BOUNDS ON THE MAXIMAL LOSS FOR THE EMPIRICAL RISK MINIMIZATION PRINCIPLE

A4.3.1 Optimistic Case

Let a set of indicator functions $Q(z, \mathbf{a}), \mathbf{a} \in \mathbf{A}$, have finite VC dimension h . In Chapter 4 we showed that for any problem π the following bound on the rate of uniform convergence is valid:

$$\begin{aligned} P \left\{ \sup_{\alpha \in \Lambda} \frac{R(\pi, \alpha) - R_{\text{emp}}(\pi, \alpha)}{\sqrt{R(\pi, \alpha)}} > \varepsilon \right\} \\ \leq \min \left(1, 4 \exp \left\{ \left(\frac{h \left(1 + \ln \frac{2\ell}{h} \right)}{\ell} - \frac{\varepsilon^2}{4} \right) \ell \right\} \right). \end{aligned} \quad (\text{A4.7})$$

When $R_{\text{emp}}(\pi, \alpha(A_{\text{emp}})) = 0$ the bound (A4.7) implies

$$\begin{aligned} P \{ R(\pi, \alpha(A_{\text{emp}})) > \varepsilon^* \} \\ \leq \min \left(1, 4 \exp \left\{ \left(\frac{h \left(1 + \ln \frac{2\ell}{h} \right)}{\ell} - \frac{\varepsilon^*}{4} \right) \ell \right\} \right) \end{aligned} \quad (\text{A4.8})$$

where we denote $\varepsilon^* = \varepsilon^2$.

Since in the optimistic case $T(\pi) = 0$ we have

$$L_M(A_{\text{emp}}) = \sup_{\pi \in \Pi} T(\pi, A_{\text{emp}}).$$

Furthermore, we have

$$\begin{aligned} T_M(A_{\text{emp}}) &= \sup_{\pi \in \Pi} \int R(\pi, \alpha(A_{\text{emp}}; z_1, \dots, z_\ell)) dF_\pi(z_1, \dots, z_\ell) \\ &= \sup_{\pi \in \Pi} \int_0^\infty P\{R(\pi, \alpha_{\text{emp}}) > \varepsilon^*\} d\varepsilon^*. \end{aligned}$$

To obtain an upper bound on this quantity we use inequality (A4.8). We obtain the inequalities

$$\begin{aligned} L_M(A_{\text{emp}}) &\leq \int_0^\infty \min \left(1, 4 \exp \left\{ \left(\frac{h \left(1 + \ln \frac{2\ell}{h} \right)}{\ell} - \frac{\varepsilon^*}{4} \right) \ell \right\} \right) d\varepsilon^* \\ &\leq \int_0^\xi d\varepsilon^* + \int_\xi^\infty 4 \exp \left\{ \left(\frac{h \left(1 + \ln \frac{2\ell}{h} \right)}{\ell} - \frac{\varepsilon^*}{4} \right) \ell \right\} d\varepsilon^* \\ &= \xi + \frac{16}{\ell} \left(\frac{2e\ell}{h} \right)^h \exp \left\{ -\frac{\xi\ell}{4} \right\}, \end{aligned} \tag{A4.9}$$

which are valid for any positive ξ . Let us choose

$$\xi = 4 \frac{h \left(\ln \frac{2\ell}{h} + 1 \right)}{\ell}, \tag{A4.10}$$

which provides a small value to the right-hand side of Eq. (A4.9). Substituting (A4.10) into (A4.9) we obtain the upper bound on the expected loss for the algorithm minimizing the empirical risk in the optimistic case:

$$L_M(A_{\text{emp}}) \leq 4 \frac{h \left(\ln \frac{2\ell}{h} + 1 \right)}{\ell} + \frac{16}{\ell}.$$

A4.3.2 Pessimistic Case

To obtain an upper bound for the pessimistic case we consider the following bound on the rate uniform convergence derived in Chapter 4:

$$\begin{aligned}
 & P \left\{ \sup_{\alpha \in \Lambda} |R(\pi, \alpha) - R_{\text{emp}}(\pi, \alpha)| > \varepsilon \right\} \\
 & \leq \min \left(1, 4 \exp \left\{ \left(\frac{h \left(1 + \ln \frac{2\ell}{h} \right)}{\ell} - \left(\varepsilon - \frac{1}{\ell} \right)^2 \right) \ell \right\} \right). \quad (\text{A4.11})
 \end{aligned}$$

This bound implies the inequality

$$\begin{aligned}
 & P \left\{ R(\pi, \alpha(A_{\text{emp}})) - R(\pi, \alpha_0) > 2\varepsilon \right\} \\
 & \leq \min \left(1, 4 \exp \left\{ \left(\frac{h \left(1 + \ln \frac{2\ell}{h} \right)}{\ell} - \left(\varepsilon - \frac{1}{\ell} \right)^2 \right) \ell \right\} \right). \quad (\text{A4.12})
 \end{aligned}$$

Indeed, from (A4.11) one obtains that with probability at least $1 - \eta$ where

$$\eta = \min \left(1, 4 \exp \left\{ \left(\frac{h \left(1 + \ln \frac{2\ell}{h} \right)}{\ell} - \left(\varepsilon - \frac{1}{\ell} \right)^2 \right) \ell \right\} \right)$$

simultaneously the following two inequalities are valid:

$$R(\pi, \alpha(A_{\text{emp}})) - R_{\text{emp}}(\pi, \alpha(A_{\text{emp}})) \leq \varepsilon$$

$$R_{\text{emp}}(\pi, \alpha_0) - R(\pi, \alpha_0) \leq \varepsilon.$$

Taking into account that

$$R_{\text{emp}}(\pi, \alpha(A_{\text{emp}})) - R_{\text{emp}}(\alpha_0) \leq 0$$

one can conclude that the inequality

$$\begin{aligned}
 & P \left\{ R(\pi, \alpha(A_{\text{emp}})) - R(\pi, \alpha_0) > 2\varepsilon \right\} \\
 & \leq \min \left(1, 4 \exp \left\{ \left(\frac{h \left(1 + \ln \frac{2\ell}{h} \right)}{\ell} - \left(\varepsilon - \frac{1}{\ell} \right)^2 \right) \ell \right\} \right)
 \end{aligned}$$

holds true.

Now let us estimate the maximal loss of ERM principle in the pessimistic

case:

$$\begin{aligned}
 L_M(A_{\text{emp}}) &= \sup_{\pi} L(\pi, A_{\text{emp}}) \\
 &= \sup_{\pi \in \Pi} \int (R(\pi, \alpha(A_{\text{emp}})) - R(\pi, \alpha_0)) dF_{\pi}(z_1, \dots, z_{\ell}) \\
 &= \sup_{\pi \in \Pi} \int_0^{\infty} P\{(R(\pi, \alpha(A_{\text{emp}})) - R(\pi, \alpha_0)) > \varepsilon\} d\varepsilon.
 \end{aligned} \tag{A4.13}$$

To get a bound we use inequality **(A4.12)**. We obtain the inequalities

$$\begin{aligned}
 L_M(A_{\text{emp}}) &\leq 2 \int_0^{\infty} \min \left(1, 4 \exp \left\{ \left(\frac{h \left(1 + \ln \frac{2\ell}{h} \right)}{\ell} - \left(\varepsilon - \frac{1}{\ell} \right)^2 \right) \ell \right\} \right) d\varepsilon \\
 &\leq 2 \left(\int_0^{\xi} d\varepsilon + 4 \int_{\xi}^{\infty} \left(\frac{2e\ell}{h} \right)^h \exp \left\{ - \left(\varepsilon - \frac{1}{\ell} \right)^2 \ell \right\} d\varepsilon \right) \\
 &= 2 \left(\xi + 4 \left(\frac{2e\ell}{h} \right)^h \int_{\xi}^{\infty} \exp \left\{ - \left(\varepsilon - \frac{1}{\ell} \right)^2 \ell \right\} d\varepsilon \right) \\
 &< 2 \left(\xi + 4 \left(\frac{2e\ell}{h} \right)^h \int_{\xi}^{\infty} \exp \left\{ - \left(\xi - \frac{1}{\ell} \right) \left(\varepsilon - \frac{1}{\ell} \right) \ell \right\} d\varepsilon \right) \\
 &= 2 \left(\xi + 4 \left(\frac{2e\ell}{h} \right)^h \frac{1}{\xi\ell - 1} \exp \left\{ - \left(\xi - \frac{1}{\ell} \right)^2 \ell \right\} \right). \tag{A4.14}
 \end{aligned}$$

This inequality is valid for any ξ . In particular, it is true for

$$\xi = \sqrt{\frac{h \left(\ln \frac{2\ell}{h} + 1 \right)}{\ell}} + \frac{1}{\ell}. \tag{A4.15}$$

Substituting **(A4.15)** into the right-hand side of Eq. (A4.14), we obtain that for $\ell > 2h$ the inequality is valid:

$$\begin{aligned}
 L_M(A_{\text{emp}}) &\leq 2 \left(\sqrt{\frac{h \left(\ln \frac{2\ell}{h} + 1 \right)}{\ell}} + \frac{4}{\sqrt{\ell h \left(\ln \frac{2\ell}{h} + 1 \right)}} \right) \\
 &< 4 \sqrt{\frac{h \left(\ln \frac{2\ell}{h} + 1 \right) + 24}{\ell}}
 \end{aligned} \tag{A4.16}$$

Thus, we obtained upper bounds on the maximal loss for the method of empirical risk minimization for both the optimistic and the pessimistic cases.

Now we would like to obtain the lower bounds on loss for the minimax loss strategy.

A4.4 LOWER BOUND FOR THE MINIMAX LOSS STRATEGY IN THE OPTIMISTIC CASE

To derive the lower bounds we use the fact that for any distribution function $P(\pi)$ the Bayesian loss does not exceed the minimax loss:

$$\inf_{A \in \mathcal{A}} L_B(A) \leq \inf_{A \in \mathcal{A}} L_M(A).$$

To estimate the lower bounds for the minimax loss we estimate the lower bound on the Bayesian loss for a special distribution function $P(\pi)$.

To construct such special distribution functions we need more detail in the description of the learning problem.

In Chapter 1, which introduced the general learning scheme, we considered three elements:

1. A generator of random vectors (it is determined by the distribution function $F(x)$)
2. A supervisor's operator $F(y|x)$ that transforms vectors x into values y
3. A set of functions of the learning machines $f(x, a)$, $a \in A$

In this setting, any specific learning problem π is determined by two elements, namely, the distribution function of the generator $F_{\pi_1}(x)$ and the supervisor's operator $F_{\pi_2}(y|x)$. Therefore to construct the distribution function on the set of problems π , one needs to consider the joint distribution function $P(\pi_1, \pi_2)$.

To obtain the Bayesian loss we consider a special distribution on the set of problems. We keep the distribution function $F(x)$ fixed for all problems and will use some distribution on the set of admissible supervisor operators.

Let a set of functions $f(x, a)$, $a \in A$ implementing a learning machine have the VC dimension h . This means that there exists h vectors such that

$$x_1, \dots, x_h \tag{A4.17}$$

can be shattered by this set of functions. Let

$$f(x, \alpha_1), \dots, f(x, \alpha_{2^h}) \tag{A4.18}$$

be the functions that shatter (A4.17).

Suppose that the probability measure $F(x)$ is concentrated on the vectors (A4.17) such that the vector x_1 has probability $1 - p$ and any other vector from (A4.17) has probability $p/(h - 1)$.

Now we define what our problems π are. In the optimistic case an admissible set of problems π (target functions) belongs to the set of functions of the learning machine. We consider 2^h different problems: Problem number π_k means to estimate the function $f(x, \alpha_k)$. To do this we are given training data

$$(x_{i_1}, f(x_{i_1}, \alpha_k)), \dots, (x_{i_\ell}, f(x_{i_\ell}, \alpha_k)) \quad (\text{A4.19})$$

containing ℓ pairs: input vectors x drawn from (A4.17) randomly and independently in accordance with the described probability measure $F(z)$ and its values $f(x, \alpha_k)$.

Assume that the a priori distribution on the set of above-mentioned problems is uniform

$$P(\pi_k) = \frac{1}{2^h}.$$

It is easy to see that in this situation the optimal Bayesian algorithm is the following: to classify vector x as $f(x, \alpha_k)$ if this vector occurs in the training set (A4.19). Classification of the vectors that do not occur in the training set (A4.19) does not matter (it can be any); that is, the optimal Bayesian algorithm for our problems is as follows: Take any function whose empirical risk is zero. The Bayesian loss for this case can be evaluated as follows:

$$\begin{aligned} \inf_{A \in \mathcal{A}} L_B(A) &= L_B(A_{\text{emp}}) \\ &= \frac{1-p}{2} p^\ell + \frac{1}{2}(h-1) \left(\frac{p}{h-1} \right) \left(1 - \frac{p}{h-1} \right)^\ell \\ &\geq \frac{p}{2} \left(1 - \frac{p}{h-1} \right)^\ell. \end{aligned} \quad (\text{A4.20})$$

Indeed in (A4.20) the value $(1-p)/2$ is the random (over all problems π) loss in classifying vector x_1 under the condition that it does not occur in the training set (A4.19); the value p^ℓ is the probability that vector x_1 does not occur in the training set (A4.19).

Analogously the value $\frac{1}{2}(p/(h-1))$ is the random loss in classifying any of the vectors x_i , $i \neq 1$; the value $(1-p/(h-1))^\ell$ is the probability that vector x_i does not occur in the training set (A4.19).

Now let us find the expression for p that maximizes the right-hand of Eq. (A4.20)

$$\frac{\partial}{\partial p} \left(1 - \frac{p}{h-1} \right)^\ell.$$

We find that

$$p = \begin{cases} 1 & \text{if } \ell \leq h-2, \\ \frac{h-1}{\ell+1} & \text{if } \ell > h-2. \end{cases} \quad (\text{A4.21})$$

Substituting (A4.21) into (A4.20), one obtains the following lower bound for the generalization ability of the empirical risk minimization principle in the

optimistic case:

$$\begin{aligned}
 L_M(A_{\text{emp}}) &\geq \inf_A L_B(A) \\
 &> \begin{cases} \frac{1}{2} \left(1 - \frac{1}{h-1}\right)^\ell & \text{if } \ell \leq h-2, \\ \frac{1}{2} \frac{h-1}{\ell+1} \left(1 - \frac{1}{\ell+1}\right)^\ell \approx \frac{1}{2e} \frac{h-1}{\ell+1} & \text{if } \ell > h-2. \end{cases} \quad (\text{A4.22})
 \end{aligned}$$

A4.5 LOWER BOUND FOR MINIMAX LOSS STRATEGY IN THE PESSIMISTIC CASE

Now we estimate the lower bound on the minimax loss for the pessimistic case. In this case using the given set of functions $f(x, a)$, $a \in A$, the learning machine tries to approximate any supervisor's rule.

As in the optimistic case we will obtain the Bayesian solution for a specific distribution functions $P(\pi)$ and will define a lower bound on the corresponding loss. This bound is a lower bound for the minimax loss.

As in the previous case we consider a learning machine that has a set of functions with VC dimension h . As before, let

$$x_1, \dots, x_h \quad (\text{A4.23})$$

be a set of vectors which can be shattered by the set of functions of the learning machine and let

$$f(x, \alpha_1), \dots, f(x, \alpha_{2^h})$$

be the functions that shatters the set of vectors (A4.23).

We will consider the following situation:

1. Probability $P(x)$ is concentrated uniformly on the vectors (A4.23)

$$P(x_i) = \frac{1}{h}$$

2. The machine solves 2^h learning problems π_k , $k = 1, \dots, 2^h$, which are determined by the following conditional probabilities

$$P_k(\omega = 0|x) = \begin{cases} 0.5 - A & \text{if } f(x, \alpha_k) = 0, \\ 0.5 + \Delta & \text{if } f(x, \alpha_k) = 1, \end{cases}$$

$$P_k(\omega = 1|x) = \begin{cases} 0.5 + A & \text{if } f(x, \alpha_k) = 0, \\ 0.5 - \Delta & \text{if } f(x, \alpha_k) = 1. \end{cases}$$

3. Probability measure on the given set of problems is uniform:

$$P(\pi_k) = \frac{1}{2^h}.$$

Under these conditions the learning machine has to select the decision rule using the training data:

$$(\omega_1, x_1), \dots, (\omega_\ell, x_\ell).$$

For any given problem π_k , the best solution of the learning machine will be the function $f(x, \alpha_k)$, which provides the smallest risk:

$$R(\pi, f(x, \alpha_k)) = \frac{1}{2} - \Delta.$$

Now our goal is to estimate Bayes' rule, which minimizes the functional

$$L_B(A) = \sum_{i=1}^{2^h} \frac{1}{2^h} \left(R(\pi, A) - \frac{1}{2} + \Delta \right).$$

The optimal algorithm for this case will be the following. Suppose that vector $z = (\omega, x)$ occurs in the training set. Let it occur $n_1(x)$ times as $(0, x)$ (as a representative of the first class) and $n_2(x)$ times as $(1, x)$ (as representative of the second class).

- If $n_1(x) > n_2(x)$, then this vector is classified as a representative of the first class.
- If on the other hand $n_1(x) < n_2(x)$, then vector x is classified as a representative of the second class.
- In the case $n_1(x) = n_2(x)$, the vector is classified as arbitrary.

If vector x does not occur in the training set, its classification does not matter (e.g., it can be done by flipping a coin).

The loss from solving any problem π_k by this algorithm is equal. Therefore

$$\min_A L_B(A) = h \left(\frac{2\Delta}{h} p_1 + \frac{\Delta}{h} p_2 \right) = 2\Delta p_1 + \Delta p_2, \quad (\text{A4.24})$$

where:

- $\frac{2\Delta}{h}$ is the loss for vector x_i , belonging to the training set, in the situation when according to function $f(x_i, \alpha_k)$ it is classified in one way and according to the described rule (using either inequality $n_1(x_i) > n_2(x_i)$ or $n_1(x_i) < n_2(x_i)$) it should be classified in the other way.

- $\frac{A}{h}$ is the loss when $n_1(x_i) = n_2(x_i)$.
- p_1 is the probability that either $n_1(x_i) < n_2(x_i)$ when $P(\omega = 0|x_i) > P(\omega = 1|x_i)$ or $n_1(x_i) > n_2(x_i)$ when $P(\omega = 0|x_i) < P(\omega = 1|x_i)$.
- p_2 is the probability of event $n_1(x_i) = n_2(x_i)$.

The exact values of p_1 and p_2 are defined by the following formulas:

$$p_1 = \sum_{\Gamma_1} \frac{\ell!}{n_1!n_2!n_3!} \left(1 - \frac{1}{h}\right)^{n_3} \left(\frac{0.5 + \Delta}{h}\right)^{n_1} \left(\frac{0.5 - \Delta}{h}\right)^{n_2}, \quad (\text{A4.25})$$

where $\Gamma_1 = \{n_1, n_2, n_3 : n_1 + n_2 + n_3 = \ell, n_1 < n_2, n_1 \geq 0, n_2 \geq 0, n_3 \geq 0\}$.

$$p_2 = \sum_{\Gamma_2} \frac{\ell!}{n_1!n_2!n_3!} \left(1 - \frac{1}{h}\right)^{n_3} \left(\frac{0.5 + \Delta}{h}\right)^{n_1} \left(\frac{0.5 - \Delta}{h}\right)^{n_2}, \quad (\text{A4.26})$$

where $\Gamma_2 = \{n_1, n_2, n_3 : n_1 + n_2 + n_3 = \ell, n_1 = n_2, n_1 \geq 0, n_2 \geq 0, n_3 \geq 0\}$.

Now we estimate the lower bound for the loss (A4.24) for different cases.

Case 1. Let $\ell \leq h$. Consider $A = 0.5$. Then using the trivial lower bounds for (A4.25) and (A4.26) (namely, $p_1 = 0$ and $p_2 \geq (1 - 1/h)^\ell$ and in accordance with (A4.24) we obtain

$$\min_A L_B(A) \geq 0.5 \left(1 - \frac{1}{h}\right)^\ell \approx 0.5 \exp\left\{-\frac{\ell}{h}\right\}.$$

Case 2. Let $h < \ell \leq 2h$. Consider $A = 0.25$. In this case for the estimation of p_1 we take into account only the term with $n_2 = 1$, and for p_2 we take into account only the term with $n_1 = n_2 = 0$. We obtain

$$\min_A L_B(A) \geq (0.25 + \frac{\ell}{8h}) \exp\left\{-\frac{\ell}{h}\right\}.$$

Case 3. Let $\ell > 2h$. Consider

$$\Delta = \frac{1}{2} \sqrt{\frac{h}{\ell}}.$$

Let us approximate the distribution of the random variable

$$\frac{n_1 - n_2}{\ell} = \theta$$

by the normal law (for definiteness sake we assume that $P(\omega = 0|x_i) > 0.5$). This random variable has the expectation

$$E\theta = \frac{2\Delta}{\ell}$$

and the variance

$$\text{Var}(\theta) = \frac{1}{\ell} \left(\frac{1}{h} - \frac{4\Delta^2}{h^2} \right) \approx \frac{1}{h\ell}.$$

Thus we consider the following normal distribution of the random variable θ :

$$P(\theta) = \frac{\sqrt{h\ell}}{\sqrt{2\pi}} \exp \left\{ -\frac{\left(\theta - \frac{2\Delta}{h} \right)^2}{2 \left(\frac{1}{\sqrt{h\ell}} \right)^2} \right\}.$$

Therefore

$$p_1 = P\{\theta < 0\} = 1 - \text{erf}\left(2A\sqrt{\frac{\ell}{h}}\right).$$

Taking into account that $A = 0.5\sqrt{\frac{h}{\ell}}$ we obtain

$$p_1 = 1 - \text{erf}(1).$$

Thus for this case

$$\min_A L_B(A) \geq \sqrt{\frac{h}{\ell}} (1 - \text{erf}(1)).$$

Combining all three cases we obtain the low bounds of minimax strategy for the pessimistic case:

$$\min_A L_B(A) \geq \begin{cases} 0.5 \exp\left\{-\frac{\ell}{h}\right\} & \text{if } \ell \leq h, \\ \left(0.25 + \frac{\ell}{8h}\right) \exp\left\{-\frac{\ell}{h}\right\} & \text{if } h < \ell \leq 2h, \\ \sqrt{\frac{h}{\ell}} (1 - \text{erf}(1)) & \text{if } \ell \geq 2h. \end{cases}$$

5

BOUNDS ON THE RISK FOR REAL-VALUED LOSS FUNCTIONS

This chapter obtains bounds on the risk for functions from a given set of real-valued functions.

We will distinguish between three cases:

1. The given set of functions is a set of totally bounded functions.
2. The given set of functions is a set of totally bounded nonnegative functions.
3. The given set of functions is a set of arbitrary nonnegative functions (it can contain unbounded functions†).

In the first and the second cases we obtain the bounds as a direct generalization of the bounds derived in Chapter 4 for sets of indicator functions.

In the third case we obtain bounds using some new concept that characterizes the tails of the distributions of a set of random variables $\xi_\alpha = Q(z, \alpha)$, $\alpha \in A$, induced by the unknown distribution function $F(z)$ and the functions in the set $Q(z, \alpha)$, $\alpha \in A$.

On the basis of these bounds, we will describe the generalization ability of minimizing the empirical risk in the set of real-valued functions.

5.1 BOUNDS FOR THE SIMPLEST MODEL: PESSIMISTIC CASE

Consider again the problem of minimizing the risk

$$R(\alpha) = \int Q(z, \alpha) dF(z) \quad (5.1)$$

†This case is important for regression estimation problems.

on the basis of empirical data

$$z_1, \dots, z_\ell,$$

where now $Q(z, a)$, $a \in A$, is a set of real-valued functions.

As before, to minimize risk (5.1) we minimize the empirical risk functional

$$R_{\text{emp}}(\alpha) = \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \quad (5.2)$$

over the set of functions $Q(z, a)$, $a \in A$.

Let the minimum of the risk (5.1) be achieved on the function $Q(z, \alpha_0)$ and the minimum of the empirical functional (5.2) on the function $Q(z, \alpha_\ell)$.

We are looking for the answers to two questions:

1. What value of risk is provided by the function $Q(z, \alpha_\ell)$? To answer this question we have to estimate the value $R(\alpha_\ell)$.
2. How close is the obtained risk to smallest possible for a given set of functions? To answer this question means to estimate the difference

$$\Delta(\alpha_\ell) = R(\alpha_\ell) - R(\alpha_0).$$

In Chapter 4 we answered these questions when $Q(z, a)$, $a \in A$, was a set of indicator functions. The goal of this chapter is to get the answers to the same questions for a set of real-valued functions.

As before, we start our study with the simplest model—that is, the case where a set of real-valued functions contains a finite number N of elements $Q(z, \alpha_k)$, $k = 1, 2, \dots, N$.

Let us estimate the rate of uniform convergence of means to their expectations over this set of functions

$$\begin{aligned} & P \left\{ \sup_{1 \leq k \leq N} \left(\int Q(z, \alpha_k) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha_k) \right) > \varepsilon \right\} \\ & \leq \sum_{k=1}^N P \left\{ \left(\int Q(z, \alpha_k) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha_k) \right) > \varepsilon \right\} \\ & \leq N \sup_{1 \leq k \leq N} P \left\{ \left(\int Q(z, \alpha_k) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha_k) \right) > \varepsilon \right\}. \end{aligned} \quad (5.3)$$

In Chapter 4, we estimated the probability of large deviations using additive Chernoff inequalities. (See Chapter 4, Eqs. (4.4) and (4.5).)

Here for the real-valued bounded function

$$A \leq Q(z, \alpha) \leq B,$$

we use Hoeffding's inequalities:

$$P \left\{ \left(\int Q(z, \alpha_k) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha_k) \right) > \varepsilon \right\} < \exp \left\{ - \frac{2\varepsilon^2 \ell}{(B-A)^2} \right\}, \quad (5.4)$$

$$P \left\{ \left(\frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha_k) - \int Q(z, \alpha_k) dF(z) \right) > \varepsilon \right\} < \exp \left\{ - \frac{2\varepsilon^2 \ell}{(B-A)^2} \right\}, \quad (5.5)$$

which are generalizations of the additive Chernoff inequalities. Using Hoeffding's inequality (5.4), we obtain from (5.3)

$$P \left\{ \sup_{1 \leq k \leq N} \left(\int Q(z, \alpha_k) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha_k) \right) > \varepsilon \right\} < N \exp \left\{ - \frac{2\varepsilon^2 \ell}{(B-A)^2} \right\}.$$

As in Chapter 4, one can rewrite this inequality in the equivalent form:

With probability $1 - \eta$ simultaneously for all N functions in the set $Q(z, \alpha_k)$, $k = 1, 2, \dots, N$, the inequality

$$\int Q(z, \alpha_k) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha_k) \leq (B - A) \sqrt{\frac{\ln N - \ln \eta}{2\ell}}$$

is valid.

Let $Q(z, \alpha_{k(0)})$ be a function from our finite set of function that minimizes the risk (5.1), and let $Q(z, \alpha_{k(\ell)})$ be a function from this set that minimizes the empirical risk (5.2). Since the obtained bound is valid simultaneously for all functions in the set, it is true as well for the function $Q(z, \alpha_{k(\ell)})$.

Thus with probability at least $1 - \eta$ the following inequality

$$\int Q(z, \alpha_{k(\ell)}) dF(z) \leq \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha_{k(\ell)}) + (B - A) \sqrt{\frac{\ln N - \ln \eta}{2\ell}} \quad (5.6)$$

is valid.

This inequality estimates the value of the risk for the chosen function $Q(z, \alpha_{k(\ell)})$. It answers the first question about estimating the risk for the function which minimizes the empirical risk in the simplest model.

To answer the second question (how close is the risk for the chosen function to the minimal one), note that for the function $Q(z, \alpha_{k(0)})$ which mini-

mizes the expected risk (5.1), Hoeffding's inequality

$$P \left\{ \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha_{k(0)}) - \int Q(z, \alpha_{k(0)}) dF(z) > \varepsilon \right\} \leq \exp \left\{ -2 \frac{\varepsilon^2 \ell}{(B - A)^2} \right\} \quad (5.7)$$

holds true.

From this inequality we find that with probability $1 - \eta$ the inequality

$$\int Q(z, \alpha_{k(0)}) dF(z) \geq \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha_{k(0)}) - (B - A) \sqrt{\frac{-\ln \eta}{2\ell}} \quad (5.8)$$

holds true.

Taking into account that $Q(z, \alpha_{k(\ell)})$ minimizes the empirical risk functional and therefore

$$\frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha_{k(0)}) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha_{k(\ell)}) \geq 0$$

from (5.6) and (5.8), we conclude that with probability at least $1 - 2\eta$ the inequality

$$\Delta(\alpha_{k(\ell)}) = R(\alpha_{k(\ell)}) - R(\alpha_{k(0)}) \leq B \sqrt{\frac{\ln N - \ln \eta}{2\ell}} + (B - A) \sqrt{\frac{-\ln \eta}{2\ell}} \quad (5.9)$$

holds true.

Thus the inequalities (5.6) and (5.9) give complete information about the generalization ability of the method of empirical risk minimization for the case when the set of totally bounded functions contains a finite number of elements: Inequality (5.6) estimates the upper bound on the risk for the chosen function, and inequality (5.9) estimates how close this bound is to the minimal possible risk for this set of functions.

These inequalities are generalizations of the analogue inequalities obtained in Chapter 4 (inequalities (4.9) and (4.12)) for a set of indicator functions.

5.2 CONCEPTS OF CAPACITY FOR THE SETS OF REAL-VALUED FUNCTIONS

5.2.1 Nonconstructive Bounds on Generalization for Sets of Real-Valued Functions

Now our goal is to generalize the results obtained for the simplest model to the general model, where the set of real-valued bounded functions $A \leq Q(z, \alpha) \leq B$, $\alpha \in A$, contains an infinite number of elements.

In Chapter 15 we prove Theorem 15.2, which, for a given probability measure, estimates the rate of uniform convergence on the set of functions

$$-\infty < A \leq Q(z, \alpha) \leq B < \infty, \quad \alpha \in \Lambda.$$

Theorem 15.2. *The inequality*

$$\begin{aligned} P \left\{ \sup_{\alpha \in \Lambda} \left| \int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \right| > \varepsilon \right\} \\ \leq \exp \left\{ \left(\frac{H_{\text{ann}}^{\Lambda}(\varepsilon/6(B-A), \ell)}{\ell} - \frac{\varepsilon^2}{36(B-A)^2} + \frac{c + \ln \ell}{\ell} \right) \ell \right\} \end{aligned}$$

is valid.

In this bound we use the concept of annealed entropy defined in Section 3.8

$$H_{\text{ann}}^{\Lambda}(E, \ell) = \ln E \ln N^{\Lambda}(E; z_1, \dots, z_\ell),$$

where $N^{\Lambda}(E; z_1, \dots, z_\ell)$ is cardinality defined in Section 3.8.

This exponential bound is nontrivial if the equality

$$\lim_{\ell \rightarrow \infty} \frac{H_{\text{ann}}^{\Lambda}(\varepsilon, \ell)}{\ell} = 0, \quad \forall \varepsilon > 0$$

is valid. In Chapter 3 we called this equality the second milestone in learning theory.

The inequality defined by Theorem 15.2 can be rewritten in the equivalent form: With probability $1 - \eta$ simultaneously for all functions the inequality

$$R(\alpha) \leq R_{\text{emp}}(\alpha) + \sqrt{\mathcal{E}(\ell)}$$

holds true, where

$$\mathcal{E}(\ell) = 36(B-A)^2 \frac{H_{\text{ann}}^{\Lambda}(\varepsilon/6(B-A), \ell) + \ln \ell + c}{\varepsilon}$$

Now one can derive the bounds on generalization ability of the machine that minimizes empirical risk in a set of real-valued functions: With probability $1 - \eta$ the inequality

$$R(\alpha_\ell) \leq R_{\text{emp}}(\alpha_\ell) + \sqrt{\mathcal{E}(\ell)}$$

holds true.

Using Hoeffding's inequality, one can obtain (exact as it was done in the last section) that with probability $1 - 2\eta$ the inequality

$$\Delta = R(\alpha_\ell) - \inf_{\alpha \in \Lambda} R(\alpha) \leq \sqrt{\mathcal{E}(\ell)} + (B-A) \sqrt{\frac{-\ln \eta}{2\ell}}$$

holds true. Therefore, using the concept of annealed entropy of sets of real-valued functions, we can construct the theory of bounds.

We, however, choose another way for developing the theory of bounds of machines minimizing the empirical risk in sets of real-valued functions. This way allows us to obtain bounds for sets of unbounded functions. The last case is important for regression estimation problem.

5.2.2 The Main Idea

In the previous chapter, construction of distribution independent bounds used a special concept of capacity of the sets of indicator functions: annealed entropy, growth function, VC dimension. Here to obtain bounds for sets of real-valued functions we generalize the capacity concept described in Chapter 4.

The idea of these generalizations is inspired by the definition of Lebesgue–Stieltjes integral. We have already used this idea, when in Chapter 2, Section 2.6 we showed that the problem of risk minimization on the basis of empirical data can be considered from the point of weak convergence of probability measures. Now we will repeat this reasoning and go a little further.

According to the definition, the Lebesgue–Stieltjes integral of a measurable nonnegative function $0 \leq \Phi(z) \leq B$ is

$$\int_0^B \Phi(z) dF(z) = \lim_{n \rightarrow \infty} \sum_{k=0}^{n-1} \frac{B}{n} P \left\{ \Phi(z) > \frac{kB}{n} \right\},$$

where

$$P \left\{ \Phi(z) > \frac{kB}{n} \right\}$$

is the probability of event

$$\mathcal{A} \left(\frac{kB}{n} \right) = \left\{ z : \Phi(z) > \frac{kB}{n} \right\}.$$

We can describe the empirical risk in a similar form:

$$\frac{1}{\ell} \sum_{i=1}^{\ell} \Phi(z_i) = \lim_{n \rightarrow \infty} \sum_{k=0}^{n-1} \frac{B}{n} \nu \left\{ \Phi(z) > \frac{kB}{n} \right\},$$

where

$$\nu \left\{ z : \Phi(z) > \frac{kB}{n} \right\}$$

is the frequency of the event $\mathcal{A}(kB/n)$ evaluated from the data z_1, \dots, z_ℓ .

Let us consider the difference

$$\begin{aligned}
 & \int_0^B \Phi(z) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} \Phi(z_i) \\
 &= \lim_{n \rightarrow \infty} \sum_{k=1}^n \frac{B}{n} P \left\{ \Phi(z) > \frac{kB}{n} \right\} - \lim_{n \rightarrow \infty} \sum_{k=1}^n \frac{B}{n} \nu \left\{ \Phi(z) > \frac{kB}{n} \right\} \\
 &= \lim_{n \rightarrow \infty} \sum_{k=1}^n \frac{B}{n} \left(P \left\{ \Phi(z) > \frac{kB}{n} \right\} - \nu \left\{ \Phi(z) > \frac{kB}{n} \right\} \right) \\
 &\leq \lim_{n \rightarrow \infty} \sum_{k=1}^n \frac{B}{n} \sup_{\beta \in (0, B)} (P \{ \Phi(z) > \beta \} - \nu \{ \Phi(z) > \beta \}) \\
 &= B \sup_{\beta \in (0, B)} (P \{ \Phi(z) > \beta \} - \nu \{ \Phi(z) > \beta \}) \\
 &= B \sup_{\beta \in (0, B)} \left(\int \theta \{ \Phi(z) - \beta \} dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} \theta \{ \Phi(z_i) - \beta \} \right),
 \end{aligned}$$

where we consider β as a parameter from the interval $(0, B)$. Let us denote this interval by B .

Thus we derived

$$\begin{aligned}
 & \int \Phi(z) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} \Phi(z_i) \\
 &\leq B \sup_{\beta \in B} \left(\int \theta \{ \Phi(z) - \beta \} dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} \theta \{ \Phi(z_i) - \beta \} \right). \quad (5.10)
 \end{aligned}$$

Below, to estimate the rate of uniform convergence for the set of bounded functions $A \subseteq Q(z, a) \leq B, a \in A$, we use the following inequality

$$\begin{aligned}
 & P \left\{ \sup_{\alpha \in \Lambda} \left(\int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \right) > \varepsilon \right\} \\
 &\leq P \left\{ \sup_{\alpha \in \Lambda, \beta \in B} \left(\int \theta \{ Q(z, \alpha) - \beta \} dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} \theta \{ Q(z_i, \alpha) - \beta \} \right) \right. \\
 &\quad \left. > \frac{\varepsilon}{B - A} \right\}. \quad (5.11)
 \end{aligned}$$

This inequality following from (5.10) is the basis for our generalizations. It shows that for any $\varepsilon > 0$ the probability that the largest deviation of averages from their expectations over a set of real-valued bounded functions $A \subseteq Q(z, a) \leq B, a \in A$, exceeds ε is less than the probability that for the set

of indicator functions $\delta\{Q(z_i, a) - \beta\}$, $a \in A$, $\beta \in D$, the largest deviation of frequencies from their probabilities exceeds $\varepsilon/(B - A)$.

In previous chapter we obtained for a set of indicator functions the bounds on the probability of the last event (bounds on the rate of uniform convergence). Using these bounds we can obtain the bounds on the rate of convergence for a set of real-valued bounded functions.

The following shows how to obtain these bounds but not before we introduce some definitions.

5.2.3 Concepts of Capacity for the Set of Real-Valued Functions

Definition of the Set of Indicators

1. Let $Q(z, a^*)$ be a real-valued function. We call the set of indicator functions

$$\theta(Q(z, \alpha^*) - \beta), \quad \beta \in \left(\inf_z Q(z, \alpha^*), \sup_z Q(z, \alpha^*) \right),$$

the *set of indicators for function $Q(z, a^*)$* (see Fig. 5.1).

2. Let $Q(z, a)$, $a \in A$, be a set of real-valued functions. We call the set of indicator functions

$$\theta(Q(z, \alpha) - \beta), \quad \alpha \in \Lambda, \beta \in \mathcal{B} = \left(\inf_{z, \alpha} Q(z, \alpha), \sup_{z, \alpha} Q(z, \alpha) \right)$$

the *complete set of indicators* for a set of real-valued functions $Q(z, a)$, $a \in A$.

Below we assume that complete set of indicators satisfies conditions of measurability for indicator functions used in Chapter 4.

Note that the set of indicators for an indicator function contains one element, namely, the indicator function. The complete set of indicators for any set of indicator functions coincides with this set of indicator functions.

According to inequality (5.11), one can obtain the bounds for the rate of uniform convergence of averages to expectations over a given set of real-valued functions by bounding the rate of uniform convergence of frequencies to probabilities over the corresponding set of indicators. We develop this idea in the following text.

We start with generalizations of the three main capacity concepts introduced in previous chapters for sets of indicator functions: the annealed entropy, the growth function, and the VC dimension.

Annealed Entropy of a Set of Indicators of Real-Valued Functions.
Let $Q(z, a)$, $z \in Z$, $a \in A$, be a set of real-valued (not necessarily bounded)

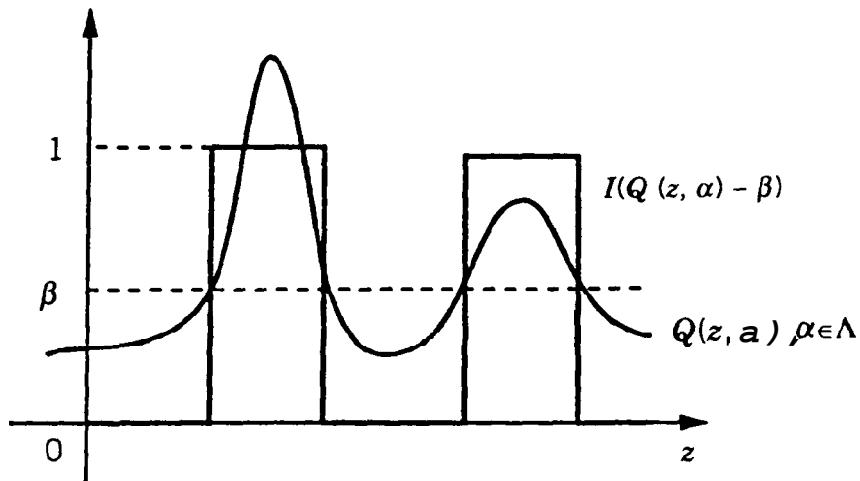


FIGURE 5.1. The indicator of level β for the function $Q(z, \alpha)$ shows for which z the function $Q(z, \alpha)$ exceeds β and for which z it does not. The function $Q(z, \alpha)$ can be described by the set of all its indicators.

functions. Let $N^{\Lambda, \beta}(z_1, \dots, z_\ell)$ be the number of *different separations* of ℓ vectors z_1, \dots, z_ℓ by a complete set of indicators:

$$\theta\{Q(z, \alpha) - \beta\}, \quad \alpha \in \Lambda, \beta \in \mathcal{B} = \left(\inf_{\alpha, z} Q(z, \alpha) \leq \beta \leq \sup_{\alpha, z} Q(z, \alpha) \right).$$

Let the function

$$H^{\Lambda, \beta}(z_1, \dots, z_\ell) = \ln N^{\Lambda, \beta}(z_1, \dots, z_\ell)$$

be measurable with respect to measure on z_1, \dots, z_ℓ .

We call the quantity

$$H_{\text{ann}}^{\Lambda, \beta}(\ell) = \ln E N^{\Lambda}(z_1, \dots, z_\ell)$$

the *annealed entropy of the set indicators of real-valued functions*.

Growth Function of a Set of Indicators of Real-Valued Function.

We call the quantity

$$G^{\Lambda, \beta}(\ell) = \ln \max_{z_1, \dots, z_\ell} N^{\Lambda, \beta}(z_1, \dots, z_\ell)$$

the *growth function of a set of real-valued functions* $Q(z, a)$, $z \in Z$, $a \in A$.

VC Dimension of a Set of Real-Valued Functions. We call the maximal number h of vectors z_1, \dots, z_h that can be shattered by the complete set of indicators $\theta\{Q(z, a) - \beta\}$, $a \in A, \beta \in \mathcal{B}$, the VC *dimension of the set of real-valued functions* $Q(z, a)$, $a \in A$.

Example. The VC dimension of a set of functions that are linear in their parameters

$$f(z, \alpha) = \sum_{i=1}^{n-1} \alpha^i \phi_i(z) + \alpha^0$$

equals n , the numbers of parameters of a set of functions.

Indeed, as was shown in Chapter 4, Section 4.11 the VC dimension of a set of linear indicator functions

$$f^*(z, \alpha) = \theta \left\{ \sum_{i=1}^{n-1} \alpha^i \phi_i(z) + \alpha^0 \right\}$$

is equal to n . The VC dimension of a set of linear functions is equal to n as well because the complete set of indicators for this set coincides with the set of linear indicator functions.

Note that all definitions are given for arbitrary sets of functions (they do not require that $\inf_{z, \alpha} Q(z, \alpha) > -\infty$ or $\sup_{z, \alpha} Q(z, \alpha) < \infty$). Note also that these definitions of the different concepts of capacity for sets of real-valued functions are generalizations of definitions of analogous concepts given in Chapters 3 and 4 for sets of indicator functions: For sets of indicator functions they coincide with the old definitions, and for sets of real-valued functions they define a new concept.

As in the case of indicator functions, these concepts are connected by the inequalities

$$H_{\text{ann}}^{\Lambda, \mathcal{B}}(\ell) \leq G^{\Lambda, \mathcal{B}}(\ell) \leq h \left(\ln \frac{\ell}{h} + 1 \right), \quad (5.12)$$

where h is the VC dimension of a set of real-valued functions $Q(z, \alpha), \alpha \in A$. Using these capacity concepts, one can obtain the bounds on uniform convergence.

5.3 BOUNDS FOR THE GENERAL MODEL: PESSIMISTIC CASE

Theorem 5.1. Let $A \leq Q(z, \alpha) \leq B$, $\alpha \in A$, be a measurable set of bounded real-valued functions, which set of indicators satisfy conditions of measurability for Theorem 4.1. Let $H_{\text{ann}}^{\Lambda, \mathcal{B}}(\ell)$ be the annealed entropy of the set of indicators.

Then the following inequality is valid:

$$\begin{aligned} & P \left\{ \sup_{\alpha \in \Lambda} \left(\int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \right) > \varepsilon \right\} \\ & \leq 4 \exp \left\{ \left(\frac{H_{\text{ann}}^{\Lambda, \mathcal{B}}(2\ell)}{\ell} - \frac{\varepsilon_*^2}{(B-A)^2} \right) \ell \right\}, \end{aligned} \quad (5.13)$$

where

$$\varepsilon_* = \varepsilon - \frac{1}{\ell}.$$

The bound obtained in this theorem to within a constant ($B - A$) coincides with the bound obtained in Theorem 4.1, this bound is nontrivial if

$$\lim_{\ell \rightarrow \infty} \frac{H_{\text{ann}}^{\Lambda, \mathcal{B}}(\ell)}{\ell} = 0.$$

The proof of this theorem is obvious: It follows from inequality (5.11) and Theorem 4.1.

Inequality (5.13) can be rewritten in the equivalent form (in the same way done several times before):

With probability $1 - \eta$ simultaneously for all functions in a set of real-valued bounded functions $Q(z, \alpha), \alpha \in \mathbf{A}$, the inequality

$$R(\alpha) \leq R_{\text{emp}}(\alpha) + (B - A)\sqrt{\mathcal{E}(\ell)} \quad (5.14)$$

is valid, where

$$\mathcal{E}(\ell) = \frac{H_{\text{ann}}^{\Lambda, \mathcal{B}}(2\ell) - \ln \eta/4}{\ell} + \frac{1}{\ell}. \quad (5.15)$$

This inequalities imply that with probability at least $1 - \eta$ the inequality

$$R(\alpha_\ell) \leq R_{\text{emp}}(\alpha_\ell) + (B - A)\sqrt{\mathcal{E}(\ell)} \quad (5.16)$$

is valid. As before, α_ℓ defines the function which yields the minimal empirical risk.

Thus we have obtained the first inequality describing the generalization ability of the learning machine minimizing the empirical risk. To derive the second inequality we use Hoeffding's inequality (5.5):

$$R(\alpha_0) > R_{\text{emp}}(\alpha_0) - (B - A)\sqrt{\frac{-\ln \eta}{2\ell}}. \quad (5.17)$$

Taking into account that for the function minimizing empirical risk with probability $1 - \eta$ inequality (5.16) holds, we conclude that with probability at least $1 - 27$ the inequality

$$\Delta(\alpha_\ell) = R(\alpha_\ell) - R(\alpha_0) \leq (B - A) \left(\sqrt{\mathcal{E}(\ell)} + \sqrt{\frac{-\ln \eta}{2\ell}} \right) \quad (5.18)$$

is valid.

The inequalities (5.16) and (5.18) describe the generalization ability of the learning machine minimizing the empirical risk in a set of totally bounded functions for a given probability measure $F(z)$.

As in the case of indicator functions, from these inequalities and inequality (5.12) one can derive both distribution-free nonconstructive bounds and distribution-free constructive bounds. To obtain these bounds, it is sufficient in the inequalities (5.16) and (5.18) to use the expression

$$\mathcal{E}(\ell) = \frac{G^{\Lambda, \mathcal{B}}(2\ell) - \ln \eta/4}{\ell} + \frac{1}{\ell}$$

(this expression provides distribution-free nonconstructive bounds), or to use the expression

$$\mathcal{E}(\ell) = \frac{h(\ln 2\ell/h + 1) - \ln \eta/4}{\ell}$$

(this expression provides distribution-free constructive bounds).

The derived bounds describe the pessimistic scenario.

5.4 THE BASIC INEQUALITY

The next sections continue to generalize the results obtained for the set of indicator functions to the set of real-valued functions.

Our goals are:

1. To obtain the bounds on the generalization ability of the empirical risk minimization induction principle for the set of real-valued bounded functions which are better than the bounds obtained in the last section.
2. To obtain the bounds on the generalization ability of the principle of empirical risk minimization for the set of real-valued unbounded functions.

We will obtain these bounds using the basic inequality to be derived in this section, which uses the auxiliary function $D_p(\alpha)$ defined as follows:

$$D_p(\alpha) = \int_0^\infty \sqrt[p]{P\{Q(z, \alpha) > c\}} dc, \quad (5.19)$$

where $1 < p \leq 2$ is some fixed parameter and $Q(z, a), a \in A$, is a set of non-negative functions.

Theorem 5.2. *Let $Q(z, a)$, $a \in A$ be a set of the real-valued (not necessary bounded) nonnegative functions. Let $H_{\text{ann}}^{\Lambda, \mathcal{B}}(\ell)$ be the annealed entropy of in-*

dicators for this set of functions. Then for any $1 < p \leq 2$ the inequality

$$P \left\{ \sup_{\alpha \in \Lambda} \frac{\int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha)}{D_p(\alpha)} > \varepsilon \right\} \\ < 4 \exp \left\{ \left(\frac{H_{\text{ann}}^{\Lambda, \mathcal{B}}(\ell)}{\ell^{2-2/p}} - \frac{\varepsilon^2}{2^{1+2/p}} \right) \ell^{2-2/p} \right\} \quad (5.20)$$

is valid.

The inequality (5.20) is nontrivial if

$$\lim_{\ell \rightarrow \infty} \frac{H_{\text{ann}}^{\Lambda, \mathcal{B}}(\ell)}{\ell^{2-2/p}} = 0.$$

Note that this theorem defines bounds for any sets of functions (not necessarily bounded).

5.4.1 Proof of Theorem 5.2

Consider the expression

$$\sup_{\alpha \in \Lambda} \frac{\int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha)}{D_p(\alpha)} \\ = \sup_{\alpha \in \Lambda} \frac{\lim_{n \rightarrow \infty} \left[\sum_{i=1}^{\infty} \frac{1}{n} P \left\{ Q(z, \alpha) > \frac{i}{n} \right\} - \sum_{i=1}^{\infty} \frac{1}{n} \nu \left\{ Q(z, \alpha) > \frac{i}{n} \right\} \right]}{D_p(\alpha)} \quad (5.21)$$

We show that if inequality

$$\sup_{\alpha \in \Lambda} \frac{P \left\{ Q(z, \alpha) > \frac{i}{n} \right\} - \nu \left\{ Q(z, \alpha) > \frac{i}{n} \right\}}{\sqrt[\ell]{P \left\{ Q(z, \alpha) > \frac{i}{n} \right\}}} \leq \varepsilon \quad (5.22)$$

is fulfilled, then the inequality

$$\sup_{\alpha \in \Lambda} \frac{\int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha)}{D_p(\alpha)} \leq \varepsilon \quad (5.23)$$

is fulfilled as well.

Indeed, (5.21) and (5.22) imply that

$$\begin{aligned} & \sup_{\alpha \in \Lambda} \frac{\int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha)}{D_p(\alpha)} \\ & \leq \sup_{\alpha \in \Lambda} \frac{\varepsilon \lim_{n \rightarrow \infty} \sum_{i=1}^{\infty} \frac{1}{n} \left(P \left\{ Q(z, \alpha) > \frac{i}{n} \right\} \right)^{1/p}}{D_p(\alpha)} = \sup_{\alpha \in \Lambda} \frac{\varepsilon D_p(\alpha)}{D_p(\alpha)} = \varepsilon. \end{aligned}$$

Therefore probability of event (5.22) does not exceed the probability of event (5.23). This means that the probability of the complementary events are connected by the inequality

$$\begin{aligned} & P \left\{ \sup_{\alpha \in \Lambda} \frac{\int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha)}{D_p(\alpha)} > \varepsilon \right\} \\ & \leq P \left\{ \sup_{\alpha \in \Lambda, \beta \in \mathcal{B}} \frac{P \{ Q(z, \alpha) > \beta \} - \nu \{ Q(z, \alpha) > \beta \}}{\sqrt[p]{P \{ Q(z, \alpha) > \beta \}}} > \varepsilon \right\}. \end{aligned}$$

In Theorem 4.2' we bounded the right-hand side of this inequality (see Chapter 4, Eq. (4.35a)). Using this bound we prove the theorem.

5.5 BOUNDS FOR THE GENERAL MODEL: UNIVERSAL CASE

Using the results of Theorem 5.2, this section derives the rate of uniform relative convergence for a bounded set of nonnegative functions $0 \leq Q(z, \alpha) \leq B$, $\alpha \in \Lambda$; that is, we prove the following theorem.

Theorem 5.3. *The inequality*

$$P \left\{ \frac{\int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha)}{\sqrt{\int Q(z, \alpha) dF(z)}} > \varepsilon \right\} \\ < 4 \exp \left\{ \left(\frac{H_{\text{ann}}^{\Lambda, \mathcal{B}}(2\ell)}{\ell} - \frac{\varepsilon^2}{4B} \right) \ell \right\} \quad (5.24)$$

is valid.

The bound (5.24) is nontrivial if

$$\lim_{\ell \rightarrow \infty} \frac{H_{\text{ann}}^{\Lambda, \mathcal{B}}(\ell)}{\ell} = 0.$$

Note also that on the right-hand side the constant B comes in first degree (rather than in squared as in Hoeffding's inequality).

Theorem 5.3 is a generalization of Theorem 4.3 obtained for a set of indicator functions.

Inequality (5.24) can be rewritten in the equivalent form:

With probability at least $1 - \eta$ simultaneously for all functions in a set of real-valued bounded functions the following inequality is fulfilled:

$$R(\alpha) \leq R_{\text{emp}}(\alpha) + \frac{B\mathcal{E}(\ell)}{2} \left(1 + \sqrt{1 + \frac{4R_{\text{emp}}(\alpha)}{B\mathcal{E}(\ell)}} \right),$$

where

$$\mathcal{E}(\ell) = 4 \frac{H_{\text{ann}}^{\Lambda, \mathcal{B}}(2\ell) - \ln \eta/4}{\ell} \quad (5.25)$$

From this inequality we find the bounds describing the generalization ability of the learning machine which minimizes the empirical risk functional:

With probability at least $1 - \eta$ the inequality

$$R(\alpha_\ell) < R_{\text{emp}}(\alpha) + B\mathcal{E}(\ell) \left(1 + \sqrt{1 + \frac{4R_{\text{emp}}(\alpha_\ell)}{\mathcal{E}(\ell)}} \right)$$

is valid, where $\mathcal{E}(\ell)$ is given by (5.25).

With probability at least $1 - 277$ the inequality

$$\Delta(\alpha_\ell) = R(\alpha_\ell) - R(\alpha_0) < B \left[\sqrt{\frac{-\ln \eta}{2\ell}} + \mathcal{E}(\ell) \left(1 + \sqrt{1 + \frac{4R_{\text{emp}}(\alpha_\ell)}{B\mathcal{E}(\ell)}} \right) \right]$$

is valid, where $\mathcal{E}(\ell)$ is given in (5.25).

These bounds are dependent on the unknown probability measure $F(z)$. As before, one obtains distribution-free nonconstructive and distribution-free constructive bounds by using the following expressions for $\mathcal{E}(\ell)$:

$$\mathcal{E}(\ell) = 4 \frac{G^{\Lambda, \mathcal{B}}(2\ell) - \ln \eta/4}{\ell},$$

$$\mathcal{E}(\ell) = 4 \frac{h \left(\ln \frac{2\ell}{h} + 1 \right) - \ln \eta/4}{\ell},$$

where $G^{\Lambda, \mathcal{B}}(\ell)$ is the growth function and h is the VC dimension of a set of real-valued functions $Q(z, \alpha), \alpha \in A$.

Theorem 5.3 completes the generalization of the theory obtained for sets of indicator functions to sets of real-valued bounded functions.

Note that when $A = 0$ and $B = 1$ the bounds on the risk for sets of bounded real-valued functions coincide with the bounds on risk for sets of indicator functions. From the conceptual point of view, the problem of minimizing the risk in sets of indicator functions (the pattern recognition problem) is equivalent to the problem of minimizing a risk in sets of real-valued bounded functions.

A new development of the problem of minimizing the risk from the data starts when one minimizes the risk in sets of nonbounded functions.

The next section analyzes this case. However, to complete this section, Theorem 5.3 must be proved.

5.5.1 Proof of Theorem 5.3

This subsection proves a more general version of Theorem 5.3.

Theorem 5.3*. *For any $1 < p \leq 2$ the inequality*

$$\begin{aligned} P \left\{ \sup_{\alpha \in \Lambda} \frac{\int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha)}{\sqrt{\int Q(z, \alpha) dF(z)}} > \varepsilon \right\} \\ < 4 \exp \left\{ \left(\frac{H_{\text{ann}}^{\Lambda, \mathcal{B}}(2\ell)}{\ell^{2(1-1/p)}} - \frac{\varepsilon^2}{2^{1+2/p} B^{2-2/p}} \right) \ell^{2(1-1/p)} \right\} \end{aligned} \quad (5.26)$$

is valid.

The proof of inequality (5.26) is based on Holder's inequality for two

functions: function[†] $f(z) \in L_p(a, b)$ and function $g(z) \in L_q(a, b)$, where

$$1/p + 1/q = 1, \quad p > 0, q > 0 :$$

$$\int_a^b |f(z)g(z)| dz \leq \left(\left(\int_a^b |f(z)|^p dz \right)^{1/p} \left(\int_a^b |g(z)|^q dz \right)^{1/q} \right)$$

Consider the function

$$D_p(\alpha) = \int_0^\infty \sqrt[p]{P\{Q(z, \alpha) > c\}} dc.$$

For a bounded set of functions we can rewrite this expression in the form

$$D_p(\alpha) = \int_0^B \sqrt[p]{P\{Q(z, \alpha) > c\}} dc.$$

Now let us denote $f(z) = \sqrt[p]{P\{Q(z, \alpha) > c\}}$ and denote $g(z) = 1$. Using these notations we utilize the Holder's inequality. We obtain

$$D_p(\alpha) = \int_0^B \sqrt[p]{P\{Q(z, \alpha) > t\}} dt < \left(\int_0^B P\{Q(z, \alpha) > t\} dt \right)^{1/p} B^{(1-1/p)}.$$

Taking into account this inequality, we obtain

$$\begin{aligned} P \left\{ \sup_{\alpha \in \Lambda} \frac{\int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha)}{\sqrt[p]{\int Q(z, \alpha) dF(z)}} > \varepsilon B^{(1-1/p)} \right\} \\ \leq P \left\{ \sup_{\alpha \in \Lambda} \frac{\int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha)}{\int \sqrt[p]{P\{Q(z, \alpha) > t\}} dt} > \varepsilon \right\}. \end{aligned}$$

Using the bound on the right-hand side of this inequality given by Theorem 5.2, we obtain inequality (5.26).

[†] Function $f(z)$ belongs to space $L_q(a, b)$ if

$$\int_a^b |f(z)|^q dz \leq \infty.$$

The values a and b are not necessarily finite.

5.6 BOUNDS FOR UNIFORM RELATIVE CONVERGENCE

This section starts the analysis of the convergence rate for sets of unbounded nonnegative functions and proves the following theorem.

Theorem 5.4. *Let the nonnegative functions (not necessary bounded) of the set $Q(z, \alpha), \alpha \in \Lambda$ be such that the random variables $\xi_\alpha = Q(z, \alpha)$ possess a finite moment of order $p > 1$. Then:*

1. *If $p > 2$, the inequality*

$$P \left\{ \sup_{\alpha \in \Lambda} \frac{\int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha)}{\sqrt[p]{\int Q^p(z, \alpha) dF(z)}} > \varepsilon a(p) \right\} \\ < 4 \exp \left\{ \left(\frac{H^{\Lambda, \mathcal{B}}(2\ell)}{\ell} - \frac{\varepsilon^2}{4} \right) \ell \right\} \quad (5.27)$$

is valid, where

$$a(p) = \sqrt[p]{\frac{1}{2} \left(\frac{p-1}{p-2} \right)^{p-1}}.$$

2. *If $1 < p \leq 2$, the inequality*

$$P \left\{ \sup_{\alpha \in \Lambda} \frac{\int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha)}{\sqrt[p]{\int Q^p(z, \alpha) dF(z)}} > \varepsilon V_p(\varepsilon) \right\} \\ < 4 \exp \left\{ \left(\frac{H^{\Lambda, \mathcal{B}}(2\ell)}{\ell^{2(1-\frac{1}{p})}} - \frac{\varepsilon^2}{2^{1+2/p}} \right) \ell^{2(1-1/p)} \right\} \quad (5.28)$$

is valid, where

$$V_p(\varepsilon) = \sqrt[p]{\left(1 - \frac{\ln \varepsilon}{\sqrt[p-1]{p}(p-1)} \right)^{p-1}}.$$

In contrast to the denominator in bound (5.26) from Theorem 5.3*, here the denominator has clear statistical meaning: It is the norm of the function $Q(z, \alpha)$ in the metric $L_p(F)$ [normalized moment of the order p of the random variable $\xi_\alpha = Q(z, \alpha)$].

Note that according to this theorem, using the normalized moment with different $p > 2$ on the left-hand side of the inequality affects only the constant in the exponent of the right-hand side of the inequality. However, the right-hand side of the inequality significantly depends on the order of the normalized moment p if $1 < p \leq 2$. This fact indicates the importance of existence of the second moment of the random variables $\xi_\alpha = Q(z, \alpha)$ for the rate of uniform convergence.

5.6.1 Proof of Theorem 5.4 for the Case $p > 2$

We prove this theorem first for $p > 2$ and then for $1 < p \leq 2$.

To prove the first part of the theorem we show that if $p > 2$ (p is not necessarily an integer), then the following inequality holds true:

$$D_2(\alpha) = \int_0^\infty \sqrt{P\{Q(z, \alpha) > c\}} dc \leq a(p) \sqrt[p]{\int Q^p(z, \alpha) dF(z)}, \quad (5.29)$$

where

$$a(p) = \sqrt[p]{\frac{1}{2} \left(\frac{p-1}{p-2} \right)^{p-1}}$$

From this inequality we find that

$$\begin{aligned} P \left\{ \sup_{\alpha \in \Lambda} \frac{\int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha)}{a(p) \sqrt[p]{\int Q^p(z, \alpha) dF(z)}} > \varepsilon \right\} \\ < P \left\{ \sup_{\alpha \in \Lambda} \frac{\int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha)}{D_2(\alpha)} > \varepsilon \right\}. \end{aligned}$$

The right-hand side of this inequality is bounded by Theorem 5.3. The first part of this theorem is the equivalent form of the inequality above. Therefore to prove the theorem it is sufficient to derive the bound (5.29).

To obtain this bound we express the functional $R(\alpha)$ in terms of the Lebesgue integral:

$$R(\alpha) = \int Q(z, \alpha) dF(z) = \int_0^\infty P\{Q(z, \alpha) > t\} dt$$

Observe that for any fixed a and arbitrary t the probability of the event $\{Q(z, a) > t\}$ can be rewritten in terms of the distribution function of the nonnegative random variable $\xi_\alpha = Q(z, a)$. Namely,

$$F_\alpha(t) = F\{\xi_\alpha \leq t\} = P\{Q(z, \alpha) \leq t\}$$

is related to the probability of the event $\{z : Q(z, a) > t\}$ as follows:

$$P\{Q(z, \alpha) > t\} = 1 - F_\alpha(t).$$

Thus the functional $R(\alpha)$ can be rewritten in the form

$$R(\alpha) = \int t dF_\alpha(t) = \int_0^\infty (1 - F_\alpha(t)) dt.$$

Moreover, the p th moment of the random variable ξ_α and the function $D_\lambda(\alpha)$ can be written as follows:

$$\begin{aligned} E\xi_\alpha^p &= \int Q^p(z, a) dF_\alpha(z) = \int t^p dF_\alpha(t) = p \prod_{t=0}^\infty t^{p-1} (1 - F_\alpha(t)) dt, \\ D_2(\alpha) &= \int_0^\infty \sqrt{P\{Q(z, \alpha) > t\}} dt = \int_0^\infty \sqrt{(1 - F_\alpha(t))} dt. \end{aligned}$$

Now let the $m_p(\alpha)$ be a moment of order $p > 2$

$$m_p(\alpha) = \int_0^\infty t^p dF_\alpha(t) = p \int_0^\infty t^{p-1} (1 - F_\alpha(t)) dt. \quad (5.30)$$

We obtain a distribution function $F_\alpha(t)$ such that $D_2(\alpha)$ is maximized for the fixed $m_p(\alpha)$.

For this purpose we construct the Lagrange function:[†]

$$\begin{aligned} L(\alpha) &= D_2(\alpha) - \mu m_p(\alpha) \\ &= \int_0^\infty \sqrt{1 - F_\alpha(t)} dt - \mu p \int_0^\infty t^{p-1} (1 - F_\alpha(t)) dt. \end{aligned} \quad (5.31)$$

We determine a probability distribution function for which the maximum of $L(\alpha)$ is obtained. Denote

$$\Phi^2 = 1 - F_\alpha(t), \quad b = \mu p$$

and rewrite (5.31) using these notations:

$$L(\alpha) = \int_0^\infty (\Phi - bt^{p-1}\Phi^2) dt$$

[†] For a review of optimization techniques using Lagrange functions see Section 9.5

The function Φ at which the maximum of the functional (5.31) is attained is defined by

$$1 - 2bpt^{p-1}\Phi = 0,$$

which implies that

$$\Phi = \left(\frac{t_0}{t} \right)^{p-1},$$

where $t_0 = (1/2pb)^{p-1}$.

Since Φ varies between 1 and 0 as t varies between 0 and ∞ , the optimal function $\Phi = \Phi(t)$ is

$$\Phi(t) = \begin{cases} 1 & \text{if } t < t_0, \\ \left(\frac{t_0}{t} \right)^{p-1} & \text{if } t \geq t_0. \end{cases}$$

We now compute $\max D_2(\alpha)$ (recalling that $p > 2$):

$$\max D_2(\alpha) = \int_0^\infty \Phi(t) dt = t_0 + \int_{t_0}^\infty \left(\frac{t_0}{t} \right)^{p-1} dt = t_0 \frac{p-1}{p-2}. \quad (5.32)$$

On the other hand, expressing $m_p(\alpha)$ in the terms of t_0 we have

$$\begin{aligned} (\text{a}) &= p \int_0^\infty t^{p-1} \Phi^2(t) dt \\ &= p \int_0^{t_0} t^{p-1} dt + \int_{t_0}^\infty t^{p-1} \Phi^{p-1}(t) dt = 2t_0^p \left(\frac{p-1}{p-2} \right). \end{aligned} \quad (5.33)$$

Substituting the value of t_0 obtained from (5.32) into (5.33), we arrive at

$$\sup_{\Phi} \frac{D_2(\alpha)}{\sqrt[p]{m_p(\alpha)}} = \sqrt[p]{\frac{1}{2} \left(\frac{p-1}{p-2} \right)^{p-1}},$$

which implies that for any $a \in A$ and $p > 2$ the inequality

$$D_2(\alpha) \leq a(p) \sqrt[p]{m_p(\alpha)}$$

holds true, where

$$a(p) = \sqrt[p]{\frac{1}{2} \left(\frac{p-1}{p-2} \right)^{p-1}}$$

Therefore the probability of event

$$\left\{ \sup_{\alpha \in A} \frac{R(\alpha) - R_{\text{emp}}(\alpha)}{\sqrt[p]{m_p(\alpha)}} > a(p)\varepsilon \right\}$$

does not exceed the probability of event

$$\left\{ \sup_{\alpha \in \Lambda} \frac{R(\alpha) - R_{\text{emp}}(\alpha)}{D_2(\alpha)} > \varepsilon \right\}.$$

According to Theorem 5.3, this fact implies the inequality

$$P \left\{ \sup_{\substack{\alpha \in \Lambda \\ \alpha \in \Lambda}} \frac{R(\alpha) - R_{\text{emp}}(\alpha)}{\sqrt{m}} > a(p)\varepsilon \right\} < \exp \left\{ \left(\frac{H^{\Lambda, B}(2\ell)}{\ell} - \frac{\alpha^2}{4} \right) \ell \right\}$$

The equivalent form of this bound is the assertion of the first part of the theorem.

5.6.2 Proof of Theorem 5.4 for the Case $1 < p \leq 2$

To prove the second part of the theorem, consider the difference

$$\begin{aligned} R(\alpha) - R_{\text{emp}}(\alpha) &= \lim_{n \rightarrow \infty} \sum_{i=1}^{\infty} \frac{1}{n} \left(P \left\{ Q(z, \alpha) > \frac{i}{n} \right\} - \nu \left\{ Q(z, \alpha) > \frac{i}{n} \right\} \right) \\ &= \int_0^{\infty} (P \{Q(z, \alpha) > t\} - \nu \{Q(z, \alpha) > t\}) dt. \end{aligned} \quad (5.34)$$

Assume that for all $a \in A$ the condition

$$R(\alpha) - R_{\text{emp}}(\alpha) \leq \varepsilon D_p(\alpha) = \varepsilon \int_0^{\infty} \sqrt[p]{P \{Q(z, \alpha) > t\}} dt \quad (5.35)$$

is fulfilled. Moreover, the inequality

$$R(\alpha) - R_{\text{emp}}(\alpha) \leq R(\alpha) = \int_0^{\infty} P \{Q(z, \alpha) > t\} dt \quad (5.36)$$

is always valid. To compute the integral (5.34) do the following: For such t that

$$P \{Q(z, a) > t\} > \varepsilon^{p/(p-1)}$$

apply the bound (5.35); otherwise, if

$$P \{Q(z, \alpha) > t\} \leq \varepsilon^{p/(p-1)},$$

apply the bound (5.36).

We thus obtain

$$\begin{aligned} R(\alpha) - R_{\text{emp}}(\alpha) \\ \leq \varepsilon \int_{1-F_\alpha(t) > \varepsilon^{p/(p-1)}} \sqrt[p]{1-F_\alpha(t)} dt + \int_{1-F_\alpha(t) \leq \varepsilon^{p/(p-1)}} (1-F_\alpha(t)) dt. \end{aligned} \quad (5.37)$$

We now find the maximal value (with respect to the distribution $F_\alpha(t)$) of the right-hand side of inequality (5.37) under the condition that the p th moment takes on some fixed value $m_p(\alpha)$; that is,

$$\int_0^\infty t^p dF_\alpha(t) = p \int_0^\infty t^{p-1} (1-F_\alpha(t)) dt = m_p(\alpha).$$

For this purpose we again use the method of Lagrange multipliers, denoting

$$\Phi^p = \Phi_\alpha^p(t) = 1 - F_\alpha(t).$$

Thus we seek the maximum of the expression

$$L(\alpha) = \int_{\Phi > \varepsilon^{1/(p-1)}} \varepsilon \Phi dt + \int_{\Phi \leq \varepsilon^{1/(p-1)}} \varepsilon \Phi^p dt - \mu \int_0^\infty t^{p-1} \Phi^p dt.$$

Represent $L(\alpha)$ in the form

$$L(\alpha) = \int_{\Phi > \varepsilon^{1/(p-1)}} (\varepsilon \Phi^p - \mu t^{p-1} \Phi^p) dt + \int_{\Phi \leq \varepsilon^{1/(p-1)}} (\varepsilon \Phi - \mu t^{p-1} \Phi^p) dt,$$

where the maximum of the first summand defines the function Φ in the domain $\Phi > \varepsilon^{1/(p-1)}$ and the maximum of the second summand in the domain $\Phi \leq \varepsilon^{1/(p-1)}$. The first summand attains its absolute maximum at

$$\Phi(t) = \sqrt[p-1]{\frac{\varepsilon}{p\mu}} \frac{1}{t}.$$

However, taking into account that Φ is a monotonically decreasing function from 1 to $\varepsilon^{1/(p-1)}$, we obtain

$$\Phi(t) = \begin{cases} 1 & \text{if } 0 \leq t < \sqrt[p-1]{\frac{\varepsilon}{p\mu}}, \\ \sqrt[p-1]{\frac{\varepsilon}{p\mu}} \frac{1}{t} & \text{if } \sqrt[p-1]{\frac{\varepsilon}{p\mu}} \leq t < \sqrt[p-1]{\frac{1}{p\mu}}. \end{cases}$$

The second summand attains its maximum in the domain $\Phi \leq \varepsilon^{1/(p-1)}$ at the function

$$\Phi(t) = \begin{cases} \sqrt[p-1]{\varepsilon} & \text{if } \sqrt[p-1]{\frac{1}{p\mu}} \leq t < \sqrt[p-1]{\frac{1}{\mu}} \\ 0 & \text{if } t > \sqrt[p-1]{\frac{1}{\mu}}. \end{cases}$$

We thus finally obtain

$$\Phi(t) = \begin{cases} 1 & \text{if } 0 \leq t < \sqrt[p-1]{\frac{\varepsilon}{p\mu}}, \\ \sqrt[p-1]{\frac{\varepsilon}{p\mu}} \frac{1}{t} & \text{if } \sqrt[p-1]{\frac{\varepsilon}{p\mu}} \leq t < \sqrt[p-1]{\frac{1}{p\mu}}, \\ \sqrt[p]{\varepsilon} & \text{if } \sqrt[p-1]{\frac{1}{p\mu}} \leq t < \sqrt[p]{\frac{1}{\mu}}, \\ 0 & \text{if } t > \sqrt[p]{\frac{1}{\mu}}. \end{cases}$$

We now express the p th moment $m_p(\alpha)$ in terms of the Lagrange multiplier μ . For this purpose we compute the p th moment:

$$m_p(\alpha) = p \int_0^\infty t^{p-1} \Phi^p(t) dt = \left(\frac{\varepsilon}{\mu} \right)^{p/(p-1)} \left(1 - \frac{\ln \varepsilon}{\sqrt[p]{p(p-1)}} \right). \quad (5.38)$$

Analogously we compute the quality

$$\begin{aligned} R(\alpha) - R_{\text{emp}}(\alpha) &\leq \varepsilon \int_0^{\sqrt[p-1]{\frac{1}{p\mu}}} \Phi(t) dt \\ &+ \int_{\sqrt[p-1]{\frac{1}{p\mu}}}^\infty \Phi^p(t) dt = \left(\frac{\varepsilon}{\mu} \right)^{1/(p-1)} \left(1 - \frac{\ln \varepsilon}{\sqrt[p]{p(p-1)}} \right). \end{aligned} \quad (5.39)$$

It follows from (5.38) and (5.39) that

$$\frac{R(\alpha) - R_{\text{emp}}(\alpha)}{\sqrt[p]{m_p(\alpha)}} \leq \varepsilon V_p(\varepsilon), \quad (5.40)$$

where we denote

$$V_p(\varepsilon) = \sqrt[p]{\left(1 - \frac{\ln \varepsilon}{\sqrt[p]{p(p-1)}} \right)^{p-1}}.$$

Thus we have shown that the condition (5.40) implies the condition (5.35). Therefore the probability of the event

$$\left\{ \sup_{\alpha \in \Lambda} \frac{R(\alpha) - R_{\text{emp}}(\alpha)}{\sqrt[p]{m_p(\alpha)}} > \varepsilon V_p(\varepsilon) \right\}$$

does not exceed the probability of event

$$\left\{ \sup_{\alpha \in \Lambda} \frac{R(\alpha) - R_{\text{emp}}(\alpha)}{D_p(\alpha)} > \varepsilon \right\}.$$

Therefore Theorem 5.2 (Eq. (5.20)) implies the inequality

$$\begin{aligned} P & \left\{ \sup_{\alpha \in \Lambda} \frac{R(\alpha) - R_{\text{emp}}(\alpha)}{\sqrt{m_p(\alpha)}} > \varepsilon V_p(\varepsilon) \right\} \\ & < 4 \exp \left\{ \left(\frac{H^{\Lambda, \mathcal{B}}(2\ell)}{\ell^{2(1-1/p)}} - \frac{\varepsilon^2}{2^{1+2/p}} \right) \ell^{2(1-1/p)} \right\}. \end{aligned}$$

The equivalent form of this bound is the assertion of the second part of our theorem.

5.7 PRIOR INFORMATION FOR THE RISK MINIMIZATION PROBLEM IN SETS OF UNBOUNDED LOSS FUNCTIONS

According to the law of large numbers, if a random variable ξ has an expectation $E\xi$, then the mean

$$\bar{\xi}_\ell = \frac{1}{\ell} \sum_{i=1}^{\ell} \xi_i$$

of ℓ independent identically distributed examples ξ_1, \dots, ξ_ℓ converges in probability to this expectation when ℓ increases.

However, the rate of convergence can be arbitrarily slow. In this case one cannot estimate the expectation using the mean $\bar{\xi}_\ell$ even if ℓ is sufficiently large.

Example. Let the random variable ξ take on the two values: 0 and K . Suppose that $P\{\xi = 0\} = 1 - \varepsilon$ and $P\{\xi = K\} = \varepsilon$ and suppose that ε is so small that with high probability $1 - \delta$ the random independent sample ξ_1, \dots, ξ_ℓ consists solely of zeros and hence the mean of this sample is zero. Probability of this event is $(1 - \varepsilon)^\ell = 1 - \delta$.

On the other hand the expectation of the random variable ξ equals $E\xi = \varepsilon K$ and, depending on value K , admits arbitrary values including the large ones (for example, when $K = l/a^2$).

Thus in our example, despite the fact that almost any training set contains only zeros, one can come to no reliable conclusions concerning the value of expectation. This happened because the distribution of the random variable was such that the "very large values" of this random variable have "sufficiently large probability," or, as some statisticians say, the distribution of random variable ξ has a "heavy tail."

To get a bound we have to possess some prior information about the "tails" of the distribution of our random variable.

From the classical statistics it is known that in order to get a bound on the means it is sufficient to know the absolute moment of any order $p > 1$ of the random variable ξ . In particular, if one knows the moment $E\xi^2$ of order $p = 2$ of the random variable ξ , then using the Chebyshev inequality one can estimate the deviation of the sum of i.i.d. values as follows:

$$P \left\{ \left| E\xi - \frac{1}{\ell} \sum_{i=1}^{\ell} \xi_i \right| > \varepsilon \right\} \leq \frac{E\xi^2}{\ell\varepsilon}.$$

In the case when one knows moment of order $1 \leq p \leq 2$, the (Barh-Essen) inequality

$$P \left\{ \left| E\xi - \frac{1}{\ell} \sum_{i=1}^{\ell} \xi_i \right| > \varepsilon \right\} < C_p \frac{E|\xi|^p}{\varepsilon^p \ell^{p-1}}$$

holds true, where $0 < C_p < 2$.

In the last section we showed that the existence of a moment of order $p > 1$ for all functions of the set $Q(z, \alpha), \alpha \in A$, with finite VC dimension implies uniform relative convergence of means to their expectations. Theorem 5.4 estimates the rate of this convergence.

However, to obtain from this rate of convergence a bound on the risk and a bound for the rate of generalization ability, it is necessary to take into account some general quantitative characteristics of admissible "tails" of distributions. In the case of a set of bounded nonnegative functions, this characteristic was B , the bound on the values of functions (the tails of the distributions of a set of random variables $\xi_\alpha = Q(z, \alpha), \alpha \in A$, is such that $P\{\xi_\alpha > B\} = 0$). The bound B is the parameter in the corresponding inequality.

In this section we consider the characteristic of the tails of a set of distributions of nonnegative random variables τ_p that depends on the parameter p , namely,

$$\sup_{\alpha \in A} \sqrt[p]{E\xi_\alpha^p} = \tau_p. \quad (5.41)$$

To show that the characteristic (5.41) describes properties of tails of distribution, let us consider a couple of examples.

Chapter 1 considers the problem of regression estimation as a problem of minimizing of the risk functional with the loss function

$$Q(z, \alpha) = (y - f(x, \alpha))^2.$$

Suppose that the distribution on the space $z = y, x$ is such that for any fixed $\alpha \in A$ the quantity $t_\alpha = y - f(x, \alpha)$ has normal distribution $N(\mu_\alpha, \sigma_\alpha^2)$ (for parameters of distribution depend on α).

Let us estimate (5.41) for $p = 2$. The quantity τ_2 is bounded by $\sqrt{3}$ (independent of parameters of distribution)

$$\begin{aligned}\tau_2 &= \frac{\sqrt{Et_\alpha^4}}{Et_\alpha^2} \\ &= \frac{\sqrt{\frac{1}{\sqrt{2\pi}\sigma_\alpha} \int_{-\infty}^{\infty} (t_\alpha)^4 \exp\left\{-\frac{(t_\alpha - \mu_\alpha)^2}{2\sigma_\alpha^2}\right\} dt_\alpha}}}{\frac{1}{\sqrt{2\pi}\sigma_\alpha} \int_{-\infty}^{\infty} (t_\alpha)^2 \exp\left\{-\frac{(t_\alpha - \mu_\alpha)^2}{2\sigma_\alpha^2}\right\} dt_\alpha} \leq \sqrt{3}.\end{aligned}$$

Indeed since

$$\frac{E(t_\alpha - \mu_\alpha)^4}{(E(t_\alpha - \mu_\alpha)^2)^2} = 3$$

the following assertion is true:

$$\begin{aligned}\tau_2 &= \frac{\sqrt{Et_\alpha^4}}{\bar{E}t_\alpha^2} = \frac{\sqrt{E((t_\alpha - \mu_\alpha) + \mu_\alpha)^4}}{E((t_\alpha - \mu_\alpha) + \mu_\alpha)^2} \\ &= \frac{\sqrt{m_4 + 6m_2\mu_\alpha^2 + \mu_\alpha^4}}{m_2 + \mu_\alpha^2} = \frac{\sqrt{3m_2^2 + 6m_2\mu_\alpha^2 + \mu_\alpha^4}}{m_2 + \mu_\alpha^2} \leq \sqrt{3},\end{aligned}$$

where we have denoted $m_4 = E(t_\alpha - \mu_\alpha)^4$ and $m_2 = E(t_\alpha - \mu_\alpha)^2$.

If the variable t_α is uniformly distributed on $(b-a, b+a)$, then taking into account that

$$\frac{E(t_\alpha - b_\alpha)^4}{(E(t_\alpha - b_\alpha)^2)^2} = \frac{9}{5}$$

one can analogously show that τ_2 has a bound:

$$\tau_2 = \frac{\sqrt{\frac{1}{2a} \int_{b-a}^{b+a} t_\alpha^4 dt_\alpha}}{\frac{1}{2a} \int_{b-a}^{b+a} t_\alpha^2 dt_\alpha} \leq \sqrt{\frac{9}{5}}.$$

Finally, if the distribution of t_α is Laplacian (double-exponential), then

$$\tau_2 = \frac{\sqrt{\frac{1}{2\Delta} \int_{-\infty}^{\infty} t_\alpha^4 \exp\left\{-\left|\frac{t_\alpha - \mu}{\Delta}\right|\right\} dt_\alpha}}{\frac{1}{2\Delta} \int_{-\infty}^{\infty} t_\alpha^2 \exp\left\{-\left|\frac{t_\alpha - \mu}{\Delta}\right|\right\} dt_\alpha} \leq \sqrt{6}.$$

Therefore τ_p describes general properties of distributions as a whole. So a priori information that $\tau_2 < 3$ means that any normal distributions, any uniform distributions, any Laplacian distribution, and many others (with "well behaved tails") are admissible.

Definition. We say that a set of nonnegative random variables $\xi_\alpha \in A$ has:

- **Distributions with light tails** if there exists a pair $(p > 2, \tau_p < c_Q)$ such that the inequality (5.41) holds true.
- **Distributions with heavy tails** if there exists a pair $(p > 1, \tau_p < c_Q)$ and there is no pair $(p > 2, \tau_p < c_Q)$ such that the inequality (5.41) holds true.

Observe that if $p < q$, then

$$\sqrt[p]{E\xi_\alpha^p} < \sqrt[q]{E\xi_\alpha^q}$$

(Liapunov inequality) is valid and consequently

$$\tau_p < \tau_q.$$

5.8 BOUNDS ON THE RISK FOR SETS OF UNBOUNDED NONNEGATIVE FUNCTIONS

Consider the set of nonnegative functions $Q(z, a), a \in A$. Suppose that the set of functions and the unknown distribution function $F(z)$ are such that corresponding distributions of the random variables $\xi_a = Q(z, a)$ have light tails. This means that we are given $p > 2$ and τ^* such that

$$\sup_{\alpha \in A} \frac{\sqrt[p]{\int Q^p(z, \alpha) dF(z)}}{\int Q(z, \alpha) dF(z)} = \tau_p < \tau^*. \quad (5.42)$$

In this case from Eq. (5.42) and the results of Theorem 5.4 for $p > 2$ one can immediately conclude that

$$\begin{aligned} P \left\{ \sup_{\alpha \in A} \frac{\int Q(z, \alpha) dF(z) - \sum_{i=1}^\ell Q(z_i, \alpha)}{\int Q(z, \alpha) dF(z)} > \tau^* a(p) \varepsilon \right\} \\ < 4 \exp \left\{ \left[\frac{H_{\text{ann}}^{\Lambda, \mathcal{B}}(2\ell)}{\ell} - \frac{\varepsilon^2}{4} \right] \ell \right\}, \end{aligned} \quad (5.43)$$

where

$$a(p) = \sqrt[p]{\frac{1}{2} \left(\frac{p-1}{p-2} \right)^{p-1}}.$$

Indeed, from the (5.42) we conclude that

$$\begin{aligned} P & \left\{ \sup_{\alpha} \frac{\int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha)}{\int Q(z, \alpha) dF(z)} > \tau^* \varepsilon \right\} \\ & < P \left\{ \sup_{\alpha} \frac{\int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha)}{\sqrt[p]{\int Q^p(z, \alpha) dF(z)}} > \varepsilon \right\} \end{aligned}$$

The right-hand side of this inequality can be bounded using the results of the Theorem 5.4.

Suppose now we face the case with heavy tails. This means that we are given $1 < p \leq 2$ and τ^* such that

$$\sup_{\alpha \in \Lambda} \frac{\sqrt[p]{\int Q^p(z, \alpha) dF(z)}}{\int Q(z, \alpha) dF(z)} = \tau_p < \tau^*. \quad (5.44)$$

In this case from Eq. (5.44) and the results of the Theorem 5.4 for the case $1 < p \leq 2$, one can analogously derive that

$$\begin{aligned} P & \left\{ \sup_{\alpha \in \Lambda} \frac{\int Q(z, \alpha) dF(z) - \sum_{i=1}^{\ell} Q(z_i, \alpha)}{\int Q(z, \alpha) dF(z)} > \tau^* \varepsilon V_p(\varepsilon) \right\} \\ & < P \left\{ \sup_{\alpha} \frac{\int Q(z, \alpha) dF(z) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha)}{\sqrt[p]{\int Q^p(z, \alpha) dF(z)}} > \varepsilon V_p(\varepsilon) \right\} \\ & < 4 \exp \left\{ \left[\frac{H_{\text{ann}}^{\Lambda, \mathcal{B}}(2\ell)}{\ell^{2(1-1/p)}} - \frac{\varepsilon^2}{2^{1+2/p}} \right] \ell^{2(1-1/p)} \right\}, \quad (5.45) \end{aligned}$$

where

$$V_p(\varepsilon) = \sqrt[p]{\left(1 - \frac{\ln \varepsilon}{\sqrt[p-1]{p}(p-1)} \right)^{p-1}}.$$

Inequalities (5.43) and (5.45) can be rewritten in equivalent form:

1. For any $p > 2$ with probability $1 - \eta$ simultaneously for all $Q(z, a)$, $a \in A$, the inequality

$$\int Q(z, a) dF(z) < \left(\frac{\frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, a)}{\frac{1 - \tau^* a(p) \sqrt{\mathcal{E}_2(\ell)}}{\ell}} \right)_{\infty}, \quad (5.46)$$

holds, where

$$\mathcal{E}_2(\ell) = 4 \frac{H_{\text{ann}}^{\Lambda, \mathcal{B}}(2\ell) - \ln \eta/4}{\ell},$$

$$\left(\frac{a}{b} \right)_{\infty} = \begin{cases} u & \text{if } b > 0, \\ \infty & \text{if } b \leq 0. \end{cases}$$

2. For any $1 < p \leq 2$ with probability $1 - \eta$ simultaneously for all $Q(z, a)$, $a \in A$, the inequality

$$\int Q(z, a) dF(z) < \left(\frac{\frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, a)}{\frac{1 - \tau^* \sqrt{\mathcal{E}_p(\ell)} V_p(\mathcal{E}_p(\ell))}{\ell^{2-2/p}}} \right)_{\infty}, \quad (5.47)$$

hold, where

$$\mathcal{E}_p(\ell) = 2^{1+2/p} \frac{H_{\text{ann}}^{\Lambda, \mathcal{B}}(2\ell) - \ln \eta}{\ell^{2-2/p}}$$

From these inequalities we find that in the case of light tails ($p > 2$) with probability $1 - \eta$ the risk for the function $Q(z, \alpha_{\ell})$, which minimizes the empirical risk, is bounded as follows:

$$\int Q(z, \alpha_{\ell}) dF(z) < \left(\frac{\frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha_{\ell})}{\frac{1 - \tau^* a(p) \sqrt{\mathcal{E}_2(\ell)}}{\ell}} \right)_{\infty}. \quad (5.48)$$

However, if the tails of distributions are heavy ($1 < p \leq 2$), then with probability $1 - \eta$ the risk is bounded by inequality

$$\int Q(z, \alpha_{\ell}) dF(z) < \left(\frac{\frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha_{\ell})}{\frac{1 - \tau^* \sqrt{\mathcal{E}_p(\ell)} V_p(\mathcal{E}_p(\ell))}{\ell^{2-2/p}}} \right)_{\infty}. \quad (5.49)$$

To get a bound on deviation $\Delta(\alpha_\ell)$, note that for $1 < q \leq 2$ the Bahr–Essen inequality

$$P \left\{ \left| EQ(z, \alpha_0) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha_0) \right| > \varepsilon \right\} < 2 \frac{EQ^q(z, \alpha_0)}{\varepsilon^q \ell^{q-1}}$$

holds true (see Section 5.7). Taking into account (5.44) and Liapunov inequality

$$\frac{\sqrt[q]{EQ^q(z, \alpha)}}{EQ(z, \alpha)} \leq \frac{\sqrt[p]{EQ^p(z, \alpha)}}{EQ(z, \alpha)}, \quad q \leq p,$$

one obtains for $1 < q \leq 2$ and $q \leq p$

$$P \left\{ \left| EQ(z, \alpha_0) - \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha_0) \right| > \varepsilon \right\} < 2 \tau^q \frac{(EQ(z, \alpha_0))^q}{\varepsilon^q \ell^{q-1}}$$

The equivalent form of this inequality is as follows: With probability at least $1 - \eta$ the inequality

$$EQ(z, \alpha_0) > \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha_0) - \frac{2^{1/q} \tau EQ(z, \alpha)}{\ell^{(1-1/q)} \sqrt[q]{\eta}} \quad (5.50)$$

is valid.

For the case $p > 2$ we have that with probability at least $1 - \eta$ the inequality

$$\Delta(\alpha_\ell) = R(\alpha_\ell) - R(\alpha_0) < \left(\frac{R_{\text{emp}}(\alpha_\ell) - R(\alpha_0) + R(\alpha_0) \tau a(p) \sqrt{\mathcal{E}_2(\ell)}}{1 - \tau a(p) \sqrt{\mathcal{E}_2(\ell)}} \right)_\infty$$

is valid. Taking into account (5.50) for $q = 2$ and inequality

$$R_{\text{emp}}(\alpha_\ell) - R_{\text{emp}}(\alpha_0) \leq 0$$

we obtain that with probability at least $1 - 277$ the inequality

$$\Delta(\alpha_\ell) \leq \tau R(\alpha_0) \left(\frac{a(p) \sqrt{\mathcal{E}_2(\ell)} + (\ell \eta)^{-1/2}}{1 - \tau a(p) \sqrt{\mathcal{E}_2(\ell)}} \right)_\infty \quad (5.51)$$

is valid. For the case $1 < p \leq 2$, choosing $q = p$ one analogously obtains

$$\Delta(\alpha_\ell) \leq \tau R(\alpha_0) \left(\frac{\sqrt{\mathcal{E}_p(\ell)} V_p(\mathcal{E}_p(\ell)) + 2^{1/p} \eta^{-1/p} \ell^{-(1-1/p)}}{1 - \tau \sqrt{\mathcal{E}_p(\ell)} V_p(\mathcal{E}_p(\ell))} \right)_\infty. \quad (5.52)$$

Therefore inequalities (5.48), (5.49), (5.51), and (5.52) describe the generalization ability of the learning machine minimizing empirical risk in a set of real-valued unbounded functions.

These inequalities contain expression $\mathcal{E}(\ell)$ that depends on the unknown distribution function.

To obtain distribution-free bounds on the generalization ability, one has to use instead of the annealed entropy $H_{\text{ann}}^{\Lambda, \mathcal{B}}(\ell)$ its upper bound obtained on the basis of the growth function $G_{\text{ann}}^{\Lambda, \mathcal{B}}(\ell)$ or the VC dimension h of a set of real-valued functions $Q(z, a), a \in A$; that is, in bounds (5.48), (5.49), (5.51), and (5.52) one has to use instead of $\mathcal{E}_2(\ell)$ the expressions

$$\mathcal{E}_2(\ell) = 4 \frac{G^{\Lambda, \mathcal{B}}(2\ell) - \ln \eta/4}{\ell},$$

$$\mathcal{E}_p(\ell) = 4 \frac{G^{\Lambda, \mathcal{B}}(2\ell) - \ln \eta/4}{\ell^{2-2/p}}$$

or the expressions

$$\mathcal{E}_2(\ell) = 4 \frac{h(\ln 2\ell/h + 1) - \ln \eta/4}{\ell},$$

$$\mathcal{E}_p(\ell) = 4 \frac{h(\ln 2\ell/h + 1) - \ln \eta/4}{\ell^{2-2/p}}$$

5.9 SAMPLE SELECTION AND THE PROBLEM OF OUTLIERS

This section discusses the idea of sample selection, which is exclusion of several elements from a given sample to determine using the remaining set, the function that yields the smallest guaranteed risk.

Note that for the pattern recognition problem the selection of a training set does not make a big difference: Minimization of the empirical risk over the entire sample, as well as doing so over a subsample of it obtained by excluding a minimal number of elements in order that the subsample could be divided without error, leads to the very same decision rule. This is a corollary of the fact that the loss function for pattern recognition takes on only two values 0 and 1. In regression problems, however, the loss function $Q(z, a)$ takes an arbitrary positive values, and therefore an exclusion of some element z may substantially change the solution as well as the estimate of the quality of the obtained solution.

Let a sample

$$z_1, \dots, z_\ell \tag{5.53}$$

be given. Consider the following

$$H(t, \ell) = \sum_{m=0}^t C_\ell^m$$

different problems of estimating the functional dependences based on empirical data

$$z_1, \dots, \hat{z}_i, \dots, \hat{z}_j, \dots, z_\ell.$$

The notation \hat{z}_i indicates that the element z_i has been excluded from (5.53). The problems differ from each other only in that for each of them the functional dependence is estimated from its own sample obtained from (5.53) by excluding at most t elements. (One can construct from (5.53) C_ℓ^m different subsamples consisting of $\ell - m$ examples. Thus altogether there are $H(t, \ell)$ different problems.)

According to the bound (5.46) for each of the $H(t, \ell)$ problems with probability $1 - \eta$ simultaneously for all $a \in A$ the inequality

$$R(\alpha) < \left(\frac{\frac{1}{\ell - t_k} \sum_{i=1}^{\ell - t_k} Q(z_i, \alpha)}{1 - \tau a(p) \sqrt{\mathcal{E}_2(\ell - t_k)}} \right)_\infty$$

holds, where t_k is the number of vectors excluded from the training data for the k th problem. Consequently, the inequalities

$$R(\alpha) < \frac{\frac{1}{\ell - t_k} \sum_{i=1}^{\ell - t_k} Q(z_i, \alpha)}{\left(1 - \tau a(p) \sqrt{\mathcal{E}_2(\ell - t_k) + \ln H(t, \ell) / (\ell - t_k)} \right)_+}$$

are valid with probability $1 - \eta$ simultaneously for all functions $Q(z, a)$, $a \in A$, in all $H(t, \ell)$ problems. Using this bound one can search for the minimum of the right-hand side over all $H(t, \ell)$ problems.

In the last formula one can use the bound on $H(t, \ell)$ derived in Chapter 4, Section 4.11:

$$\ln H(t, \ell) \leq t \left(\ln \frac{\ell}{t} + 1 \right).$$

Thus, in searching for the best guaranteed solutions using empirical risk minimization method, one can try to exclude some subset of training data to obtain the best bound.

The excluded data (which cause a decrease in the guaranteed minimum of risk) can be called outliers.

5.10 THE MAIN RESULTS OF THE THEORY OF BOUNDS

In this chapter we obtained the main bounds describing the generalization ability of the learning machines.

To obtain these bounds we introduced several general (which are valid for any set of functions) concepts of capacity.

We showed that one has to distinguish between two different cases of the problem of risk minimization from empirical data: the case where the admissible set of functions is a set of *totally bounded functions* (we considered the case $0 \leq Q(z, a) \leq B$, $a \in A$) and the case where the admissible set contains *unbounded functions*.

In the first case one obtains the same type of bounds as in pattern recognition: namely, with probability at least $1 - \eta$ the bound

$$R(\alpha_\ell) < R_{\text{emp}}(\alpha_\ell) + \frac{B\mathcal{E}(\ell)}{2} \left(1 + \sqrt{1 + \frac{4R_{\text{emp}}(\alpha_\ell)}{B\mathcal{E}(\ell)}} \right)$$

is valid and with probability at least $1 - 2\eta$ the bound

$$\Delta(\alpha_\ell) \leq R_{\text{emp}}(\alpha_\ell) + \frac{B\mathcal{E}(\ell)}{2} \left(1 + \sqrt{1 + \frac{4R_{\text{emp}}(\alpha_\ell)}{B\mathcal{E}(\ell)}} \right) + \sqrt{\frac{-\ln \eta}{\ell}}$$

is valid, where $\mathcal{E}(\ell)$ to within a small value is ratio of capacity function over number of observations

$$\mathcal{E}(\ell) \approx \frac{\text{capacity characteristic}}{\ell} \quad (5.54)$$

In (5.54) one can use any capacity function determined in this chapter.

In the second case we obtained bounds for nonnegative loss functions using a priori information about the tails of distributions that concentrate in the pair[†] ($p > 1$, and τ_p):

$$\sup_{\alpha, z} \frac{\sqrt[p]{EQ^p(z, \alpha)}}{EQ(z, \alpha)} \leq \tau_p$$

Knowing the value τ_p , we derived that with probability at least $1 - \eta$ the inequality

$$R(\alpha_\ell) \leq \left(\frac{R_{\text{emp}}(\alpha_\ell)}{1 - \tau a(p) \sqrt{\mathcal{E}_2(\ell)}} \right)_\infty, \quad a(p) = \sqrt[p]{\frac{1}{2} \left(\frac{p-1}{p-2} \right)^{p-1}}$$

[†]Here we consider the case $p > 2$ only to simplify formulas.

is valid and with probability at least $1 - 2\eta$ the inequality

$$\Delta(\alpha_\ell) < \tau R(\alpha_0) \left(\frac{a(p)\sqrt{\mathcal{E}(\ell)} + (\eta\ell)^{-1/2}}{1 - \tau a(p)\sqrt{\mathcal{E}(\ell)}} \right)_\infty$$

is valid, where $\mathcal{E}(\ell)$ is determined by expression of type (5.54).

Expression (5.54) is one of the most important points of the theory. In the extreme case when one knows the probability measure $F(z)$ and can evaluate the annealed entropy, one obtains the distribution-specific bounds using

$$\mathcal{E}(\ell) = 4 \frac{H_{\text{ann}}^{\Lambda, \mathcal{B}}(2\ell) - \ln \eta/4}{\ell}$$

In another extreme case when no a priori information is found about the probability measure, the distribution-free bounds are determined on the basis of the growth function $G^{\Lambda, \mathcal{B}}(\ell)$ of a set of real-valued functions $Q(z, a)$, $a \in \mathbf{A}$:

$$\mathcal{E}(\ell) = 4 \frac{G^{\Lambda, \mathcal{B}}(2\ell) - \ln \eta/4}{\ell}$$

To obtain constructive distribution-free bounds we found the upper bound for the growth function using the VC dimension concept:

$$\mathcal{E}(\ell) \leq 4 \frac{\ln 2\ell/h + 1}{\ell/h} + \frac{-\ln \eta/4}{\ell}$$

Moreover, the theory shows a clear way how to construct rigorous distribution-dependent bounds. To get nonconstructive rigorous bounds, one has to use the expression

$$\mathcal{E}(\ell) = 4 \frac{M^{\Lambda, \mathcal{B}}(2\ell) - \ln \eta/4}{\ell}$$

with the generalized growth function

$$M^{\Lambda, \mathcal{B}}(\ell) = \sup_{P \in \mathcal{P}} \ln E_P N^{\Lambda, \mathcal{B}}(z_1, \dots, z_\ell).$$

To make the bounds constructive, one has to find a way to obtain the bound for the generalized growth function that is better than one based on the VC dimension.

[†] It is remarkable that to within a small value this functional form depends on the ratio ℓ/h of the number of observations over the VC dimension of the set of functions.

6

THE STRUCTURAL RISK MINIMIZATION PRINCIPLE

This chapter addresses methods for controlling the generalization ability of learning machines that use small size samples of training instances.

We consider the sample of size ℓ to be small if the ratio ℓ/h (ratio of the number of the training patterns to the VC dimension of the set of functions of the learning machines) is small, say $\ell/h < 20$.

The induction principle for learning from samples of small size, the so-called Structural Risk Minimization (SRM) principle is introduced first. In contrast to the Empirical Risk Minimization principle, which suggests that we should minimize the empirical risk at any cost, this principle looks for the optimal relationship between the amount of empirical data, the quality of approximation of the data by the function chosen from a given set of functions, and the value that characterizes capacity of a set of functions. The SRM principle finds the function that for the fixed amount of data achieves the minimum of the guaranteed risk.

In the case of the pattern recognition problem, we compare the SRM principle to another small sample size induction principle, namely, the so-called Minimum Description Length (MDL) principle.

Then we show that the SRM method is always consistent and we derive a bound on the asymptotic rate of convergence.

At the end of the chapter we consider the problem of minimizing the Local Risk Functional, whose solution is based on the SRM principle.

6.1 THE SCHEME OF THE STRUCTURAL RISK MINIMIZATION INDUCTION PRINCIPLE

In the last two chapters we obtained the bounds on the risk which are valid simultaneously for all functions in a given set of functions.

We proved that with probability at least $1 - \eta$ simultaneously for all functions from the set of totally bounded functions $0 \leq Q(z, a) \leq B$, $a \in \mathbf{A}$, with finite **VC** dimension h the (additive) inequality

$$R(\alpha) \leq R_{\text{emp}}(\alpha) + \frac{B\mathcal{E}(\ell)}{2} \left(1 + \sqrt{1 + \frac{4R_{\text{emp}}(\alpha)}{B\mathcal{E}(\ell)}} \right) \quad (6.1)$$

holds true,[†] where

$$\mathcal{E}(\ell) = 4 \frac{h \left(\ln \frac{2\ell}{h} + 1 \right) - \ln \eta/4}{\ell}. \quad (6.2)$$

We proved also that if a pair (p, τ) exists such that for all functions from the set of nonnegative (not necessarily bounded) functions $Q(z, a)$, $a \in \mathbf{A}$, with the **VC** dimension h the inequality

$$\frac{\sqrt[p]{EQ^p(x, \alpha)}}{EQ(z, \alpha)} \leq \tau, \quad p > 2 \quad (6.3)$$

holds true (the corresponding set of random variables contains only light tails), then with probability at least $1 - \eta$ simultaneously for all functions $Q(z, a)$, $a \in \mathbf{A}$, the (multiplicative) inequality

$$\begin{aligned} R(\alpha) &\leq \left(\frac{R_{\text{emp}}(\alpha)}{1 - a(p)\tau\sqrt{\mathcal{E}(\ell)}} \right)_\infty, \\ a(p) &= \sqrt[p]{\frac{1}{2} \left(\frac{p-1}{p-2} \right)^{p-1}}, \end{aligned} \quad (6.4)$$

holds.[‡]

Now using these inequalities we would like to control the process of minimizing the risk functional on the basis of fixed amount of empirical data.

The simplest way to control this process is to minimize the value of empirical risk. According to inequalities (6.1) and (6.4) the upper bound on the risk decreases with decreasing the value of empirical risk. This is the reason why the principle of empirical risk minimization often gives good results for *large sample size*. If it happens that ℓ/h is large, then the value of actual risk is determined by the value of empirical risk. Therefore to minimize actual risk one minimizes the empirical risk.

However, if ℓ/h is small, a small value of empirical risk $R_{\text{emp}}(\alpha_\ell)$ does not guarantee a small value of the actual risk. In this case, to minimize the actual risk $R(\alpha)$ one has to minimize the right-hand side of inequality (6.1) (or (6.4))

[†]To control the generalization ability of learning machines we need constructive bounds on the risk. Therefore in this chapter we will use distribution free constructive bounds.

[‡]Here only for simplicity of notation we consider the case $p > 2$. The case $1 < p \leq 2$ can be considered as well.

simultaneously over both terms. Note that the first term in inequality (6.1) depends on a specific function of the set of functions, while for a fixed number of observations the second term depends mainly on the VC dimension of the whole set of functions.

Therefore to minimize the right-hand side of the bound of risk, (6.1) (or (6.4)), simultaneously over both terms, one has to make the VC dimension a controlling variable.

To do this we consider the following scheme.

6.1.1 Principle of Structural Risk Minimization

Let us impose the structure \mathbf{S} on the set S of functions $Q(z, a)$, $a \in A$, with a structure \mathbf{S} . Consider the set of nested subsets of functions (Fig. 6.1)

$$S_1 \subset S_2 \subset \cdots \subset S_n, \dots, \quad (6.5)$$

where $S_k = \{Q(z, a) : a \in \Lambda_k\}$, and

$$S^* = \bigcup_k S_k.$$

Consider admissible structures—the structures that satisfy the following properties:

1. Any element S_k of structure \mathbf{S} has a finite VC dimension h_k .
 2. Any element S_k of the structure (6.5) contains either
- (i) a set of totally bounded functions

$$0 \leq Q(z, \alpha) \leq B_k, \quad \alpha \in \Lambda_k$$

- (ii) or a set of nonnegative functions $Q(z, a)$, $a \in \Lambda_k$, satisfying the inequality

$$\sup_{\alpha \in \Lambda_k} \frac{\sqrt[p]{EQ^p(z, \alpha)}}{EQ(z, \alpha)} \leq \tau_k < \infty. \quad (6.6)$$

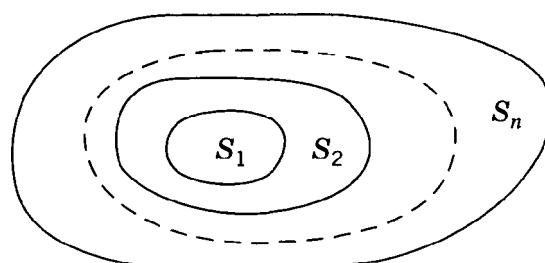


FIGURE 6.1. A structure on the set of functions is determined by the nested subsets of functions.

3. The set S^* is everywhere dense in the set S in the $L_1(F)$ metric[†] where $F = F(z)$ is the distribution function from which examples are drawn.

Note that in view of (6.5) the following assertions are true:

1. The sequence of values of VC dimensions h_k for the elements S_k of the structure \mathbf{S} is nondecreasing with increasing k

$$h_1 \leq h_2 \leq \dots < h_n \leq \dots$$

- 2a. The sequence of values of the bounds B_k for the elements S_k of the structure \mathbf{S} is nondecreasing with increasing k :

$$B_1 \leq B_2 \leq \dots \leq B_n \leq \dots \quad (6.7)$$

- 2b. The sequence of values of the bounds τ_k for the elements S_k of the structure \mathcal{S} is nondecreasing with increasing k

$$\tau_1 \leq \tau_2 \leq \dots \leq \tau_n \leq \dots$$

Denote by $Q(z, \alpha_\ell^k)$ the function that minimizes the empirical risk in the set of functions S_k . Then with probability $1 - \eta$ one can assert that the actual risk for this function is bounded as

$$R(\alpha_\ell^k) \leq R_{\text{emp}}(\alpha_\ell^k) + B_k \mathcal{E}_k(\ell) \left(1 + \sqrt{1 + \frac{4R_{\text{emp}}(\alpha_\ell^k)}{B_k \mathcal{E}_k(\ell)}} \right), \quad (6.8)$$

or as

$$R(\alpha_\ell^k) \leq \left(\frac{R_{\text{emp}}(\alpha_\ell^k)}{1 - a(p)\tau_k \sqrt{\mathcal{E}_k(\ell)}} \right)_\infty, \quad (6.9)$$

where

$$\mathcal{E}_k(\ell) = 4 \frac{h_k \left(\ln \frac{2\ell}{h_k} + 1 \right) - \ln \eta / 4}{\ell} \quad (6.10)$$

[†] We will need this property for asymptotic analysis of SRM principle, when structure contains an infinite number of elements.

The set S^* is everywhere dense in the set S in $L_1(F)$ metric if for any $\delta > 0$ and any function $Q(z, \alpha) \in S$ there exists a function $Q(z, \alpha^*) \in S^*$ such that

$$\rho(Q(z, \alpha), Q(z, \alpha^*)) = \int |Q(z, \alpha) - Q(z, \alpha^*)| dF(z) < \delta.$$

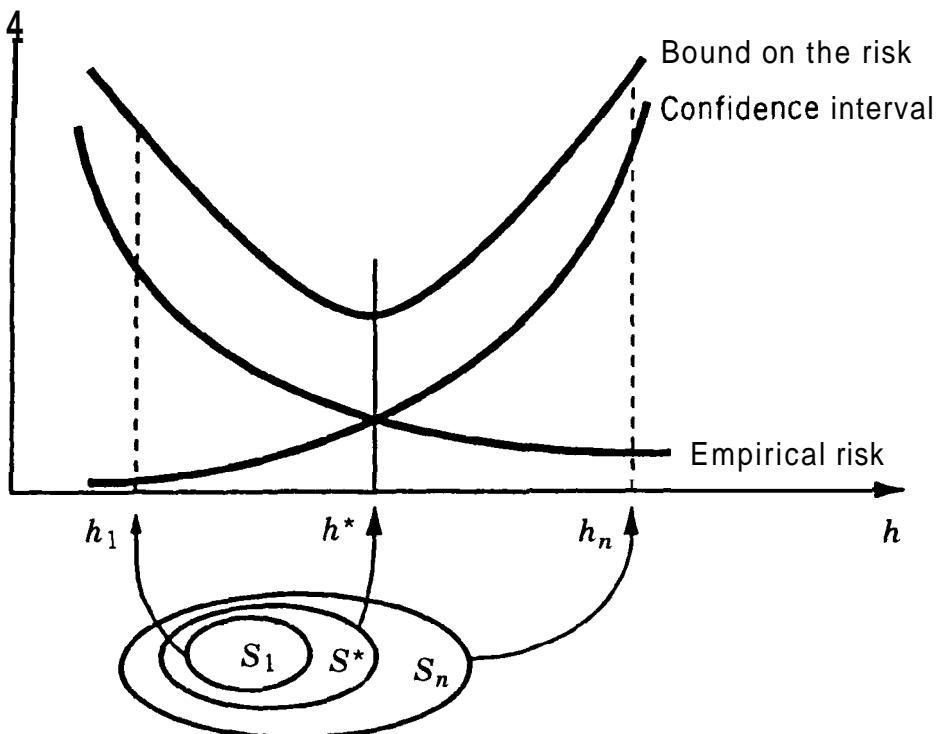


FIGURE 6.2. The bound on the risk is the sum of the empirical risk and of the confidence interval. The empirical risk is decreased with the index of element of the structure, while the confidence interval is increased. The smallest bound of the risk is achieved on some appropriate element of the structure.

For a given set of observations z_1, \dots, z_ℓ , the SRM method chooses the element S_k of the structure for which the smallest bound on the risk (the smallest guaranteed risk) is achieved.

Therefore the idea of the structural risk minimization induction principle is the following:

To provide the given set of functions with an admissible structure and then to find the function that minimizes guaranteed risk (6.8) (or (6.9)) over given elements of the structure.

To stress the importance of choosing the element of the structure that possesses an appropriate capacity, we call this principle the principle of structural risk minimization. It describes a general model of capacity control. To find the guaranteed risk, one has to use bounds on the actual risk. As shown in previous chapters, all of them have to contain information about the capacity of the element of the structure to which the chosen function belongs. In this chapter, we will use the bounds (6.8) or (6.9).

Section 6.3 shows that the SRM principle is always consistent and defines a bound on the rate of convergence. However, we must first describe the Minimum Description Length principle and point out its connection to the SRM principle for pattern recognition problem.

6.2 MINIMUM DESCRIPTION LENGTH AND STRUCTURAL RISK MINIMIZATION INDUCTIVE PRINCIPLES

6.2.1 The Idea About the Nature of Random Phenomena

In the 1930s Kolmogorov introduced the axioms of probability theory. Subsequently, probability theory became a purely mathematical (i.e., deductive) science. This means that developing the theory became the result of formal inference, based on some rules of inference. Axiomatization of the theory, however, removed from consideration a very important question, namely, one about the *nature of randomness*. The theory of probability does not answer the question, *What is randomness?* It simply ignores it by using the axioms about *given probability measures* (see Chapter 2). Nevertheless, the question remains and needs to be answered.

Thirty years after axiomatization of the probability theory Solomonoff (1960), Kolmogorov (1965) and Chaitin (1966) suggested the model of randomness. This model was constructed on the basis of a new concept, the so-called *algorithmic (descriptive) complexity*.

The algorithmic complexity on the object is defined to be the length of the shortest binary computer program that describes this object. It was proved that the value of algorithmic complexity up to an additive constant does not depend on the type of computer. Therefore it is a universal characteristic of the object.

Now one can compare the given length of object description with its algorithmic complexity. The main idea is as follows:

To consider a relatively large string describing an object to be random if algorithmic complexity of an object is high—that is, if the given description of an object cannot be compressed significantly.

Shortly after the concept of algorithmic complexity was introduced, first Wallace and Boulton (1968) and then Rissanen (1978) suggested that we use the concept of algorithmic complexity as a main tool of induction inference of learning machines; they suggest an induction principle that was called the Minimum Message Length (MML) principle by Wallace and Boulton, and the Minimum Description Length (MDL) principle by Rissanen.

6.2.2 Minimum Description Length Principle for the Pattern Recognition Problem

Suppose we are given training data. That is, we are given ℓ pairs containing the vector x and the binary value w

$$(\omega_1, x_1), \dots, (\omega_\ell, x_\ell) \quad (6.11)$$

(pairs drawn randomly and independently according to some probability measure). Consider two strings: the binary string

$$\omega_1, \dots, \omega_\ell \quad (6.12)$$

and the string of vectors

$$x_1, \dots, x_f. \quad (6.13)$$

The question is, *Given (6.13), is the string (6.12) a random object?* To answer this question let us analyze the complexity of the string (6.12) in the spirit of Solomonoff–Kolmogorov–Chaitin ideas. Since ω_i , $i = 1, \dots, \ell$ are binary values, the string (6.12) is described by ℓ bits.

To determine the complexity of this string, let us try to compress its description. Since training pairs were drawn randomly and independently, the value ω_i may depend only on the vector x_i but not on the vector x_j , $i \neq j$ (of course, only if the dependency exists).

Consider the following model: We are given a fixed code book C_b with $N \ll 2^\ell$ different tables T_i , $i = 1, \dots, N$. Any table T_i describes some function+from x to ω .

Try to find in the code book C_b the table T that describes the string (6.12) in the best possible way; namely, the table on which the given string (6.13) returns the binary string

$$\omega_1^*, \dots, \omega_\ell^* \quad (6.14)$$

such that the Hamming distance between strings (6.12) and (6.14) is minimal (the number of errors in decoding (6.12) by this table T is minimal).

Suppose we have found a perfect table T_o for which the Hamming distance between (6.14) and (6.12) is zero.

This table decodes the string (6.12).

Since the code book C_b is fixed, to describe the string (6.12) it is enough to specify the number o of table T_o in the code book. If there is no a priori information about the desired table, then the minimal number of bits needed to decode the number of one of the N tables is $\lceil \log_2 N \rceil$, where $\lceil A \rceil$ is the minimal integer no less than A . Therefore in this case to describe (6.12) we need $\lceil \log_2 N \rceil$ bits, rather than ℓ . Thus, using a code book with a perfect decoding table, we compress the description length of string (6.12)

$$K(T_o) = \frac{\lceil \log_2 N \rceil}{\ell} \quad (6.15)$$

times. Let us call $K(T)$ the coefficient of compression in the description of the string (6.12).

[†]Formally speaking, to have the finite tables in the code book, the input vector x has to be discrete. However, as we will see, the number of levels of quantization will not affect the bounds on risk. Therefore one can consider any degree of quantization, even the tables with infinite number of entries.

Now consider the general case: Code book C_b does not contain the perfect table. Let the smallest Hamming distance between strings (obtained and desired) be $d \geq 0$. Without loss of generality one can assume $d < \ell/2$. Otherwise, instead of the smallest distance, one will look for the largest Hamming distance and during decoding change 1 to 0 and vice versa. This will cost one extra bit in the coding scheme.

For fixed d there are C_ℓ^d different possible corrections to the string of length ℓ . To specify one of them (i.e., to specify a number of one of the C_ℓ^d variants) one needs $\lceil \log_2 C_\ell^d \rceil$ bits.

Therefore to describe string (6.12) we need $\lceil \log_2 N \rceil$ bits to describe the number of the table and we need $\lceil \log_2 C_\ell^d \rceil$ bits to describe the number of correction variant. All together we need $\lceil \log_2 N \rceil + \lceil \log_2 C_\ell^d \rceil$ bits for describing (6.12). If d is unknown, we need additional Δ bits to define it. In this case our description contains

$$\lceil \log_2 N \rceil + \lceil \log_2 C_\ell^d \rceil + \Delta$$

bits information. This number should be compared to ℓ , the number of bits in the description of the string (6.11). Therefore the coefficient of compression is

$$K(T) = \frac{\lceil \log_2 N \rceil + \lceil \log_2 C_\ell^d \rceil + \Delta}{\ell} \quad (6.16)$$

If the coefficient of compression $K(T)$ (or $K_o(T)$) is small, then according to the Solomonoff–Kolmogorov–Chaitin idea the string is not random and somehow depends on the input vectors \mathbf{x} . The decoding table T somehow approximates the unknown functional relation between \mathbf{x} and ω .

6.2.3 Bounds for the Minimum Description Length Principle

The question is, *Does the compression coefficient $K(T)$ determine the probability of the test error in classification (decoding) vectors x by the table T ?* The answer is yes.

To prove this, compare the result obtained for the MDL principle to the result obtained for ERM principle in the simplest model.

In Chapter 4 Section 4.3 we obtained the bound that if a set of functions contains N elements then with probability at least $1 - \eta$ the inequality

$$R(T_i) \leq R_{\text{emp}}(T_i) + \frac{\ln N - \ln \eta}{\ell} \left(1 + \sqrt{1 + \frac{2R_{\text{emp}}(T_i)\ell}{\ln N - \ln \eta}} \right) \quad (6.17)$$

holds true simultaneously for all N functions in the given set of functions (for all N tables in the given code book). Let us transform the right-hand side of this inequality using concept of compression coefficient and the fact

that

$$R_{\text{emp}}(T_i) = \frac{d}{\ell}.$$

Note that for $d \leq \ell/2$, $\ell > 6$, and $A \geq 0$ the inequality

$$\begin{aligned} & \frac{d}{\ell} + \frac{\ln N - \ln \eta}{\ell} \left(1 + \sqrt{1 + \frac{2d}{\ln N - \ln \eta}} \right) \\ & < 2 \left(\frac{\lceil \ln N \rceil + \lceil \ln C_\ell^d \rceil + \Delta}{\ell} - \frac{\ln \eta}{\ell} \right) \end{aligned} \quad (6.18)$$

is valid (one can easily check it). Now rewrite the right-hand side of (6.17) in terms of the compression coefficient (6.16)

$$2 \ln 2 \left(\frac{\lceil \log_2 N \rceil + \lceil \log_2 C_\ell^d \rceil + \Delta}{\ell} - \frac{\log_2 \eta}{\ell} \right) = 2 \left(\ln 2 K(T) - \frac{\ln \eta}{\ell} \right).$$

Since inequality (6.17) holds true with probability at least $1 - \eta$ and inequality (6.18) holds with probability 1, then the inequality

$$R(T_i) < 2 \left(\ln 2 K(T_i) - \frac{\ln \eta}{\ell} \right) \quad (6.19)$$

holds with probability at least $1 - \eta$.

6.2.4 Structural Risk Minimization for the Simplest Model and Minimum Description Length Principle

Now suppose that we are given $M \ll 2^\ell$ code books that make up a structure: code book 1 contains a small number of tables, code book 2 contains these tables and some additional tables, and so on.

Now describe the string (6.12) using a more sophisticated decoding scheme: First describe the number m of the code book and then using this code book describe the string (as we showed above it takes $\lceil \log_2 N_m \rceil + \lceil \log_2 C_\ell^d \rceil$ bits, where N_m is the number of tables in the m th code book).

The total length of description in this case is no less than $\lceil \log_2 N_m \rceil + \lceil \log_2 C_\ell^d \rceil$ and the compression coefficient is not less than

$$K(T) \leq \frac{\lceil \log_2 N_m \rceil + \lceil \log_2 C_\ell^d \rceil + \Delta + \log_2 m}{\ell}$$

For this case the inequality (6.18) holds. Therefore the probability of error for the table which was used for compressing the description of string (6.12) is bounded by (6.19).

Thus we have proven the following theorem.

Theorem 6.1. *If on the given structure of code books one compresses $K(T)$ times the description of string (6.12) using a table T, then for $\ell > 6 \ln d < \ell/2$ with probability at least $1 - \eta$ one can assert that the probability of committing an error by the table T is bounded as follows:*

$$R(T) < 2 \left(\log 2K(T) - \frac{\ln \eta/m}{\ell} \right). \quad (6.20)$$

Note how powerful the concept of compression coefficient is: To obtain bound for the probability of error we actually need only information about this coefficient.[†] We are not interested in such details as:

How many examples we used.

How the structure of code books was organized.

Which code book was used and how many tables were in this code book.

How many errors were made by the table from the code book we used.

Nevertheless, the value of bound (6.20) does not exceed very much the value of the bound of the risk (6.17) obtained on the basis of the theory of uniform convergence, which has a more sophisticated structure and which uses information about the number of functions (tables) in the sets, number of errors in the training set, and number of elements of the training set.

Note also that within a factor of 2 the bound (6.20) cannot be improved: In the case when a perfect table exists in the code book, equality can be achieved with the factor of 1.

This theorem justifies the MDL principle: To minimize the probability of error one has to minimize the coefficient of compression.

6.2.5 The Shortcoming of the Minimum Description Length Principle

There exists, however, a shortcoming of the MDL principle. Recall that the MDL method uses code books with a *finite number* of tables. Therefore, in order to deal with a set of functions that continuously depends on parameters, one has to first quantize that set to make the tables.

Quantization can be done in many ways. The problem is, *How do we make the "smart" quantization for a given number of observations?* For a given set of functions, how can we construct a code book with a small number of tables but with good approximation ability?

[†]For not very large M (say $M < \ell^k$, $k \ll \log_2 N_m$) the second term $(\ln M - \ln \eta)/\ell$ on the right-hand side is actually foolproof: For reasonable η and ℓ , it is small compared to the first term, but it prevents us from considering too small η or/and too small ℓ .

A good quantization essentially can reduce the number of tables in the code book, effecting the compression coefficient. Unfortunately, finding a good quantization is extremely difficult and determines the main shortcoming of MDL principle.

Chapter 10 constructs a set of linear functions in very high-dimensional space (experiments described in Chapter 12, use linear functions in $N \sim 10^{13}$ -dimensional space) that has low VC dimension (in these experiments, $h \sim 10^2 - 10^3$). One can guarantee that if a function from this set separates a training set of size ℓ without error, then the probability of test error is proportional to $h \ln \ell / \ell$.

The problem for the MDL approach to this set of indicator functions is, *How do we construct code books with $\sim \ell^h$ tables (but not with $\sim \ell^N$ tables) that approximate this set of linear functions well?*

The MDL principle works well when the problem of constructing reasonable code books has a good solution.

6.3 CONSISTENCY OF THE STRUCTURAL RISK MINIMIZATION PRINCIPLE AND ASYMPTOTIC BOUNDS ON THE RATE OF CONVERGENCE

Let us continue the study of the SRM principle. In this section we analyze asymptotic properties of the SRM principle. Here we answer two questions:

1. Is the Structural Risk Minimization principle consistent? (Do the risks for the functions chosen according to this principle converge to the smallest possible risk for the set S with increasing amount of observations?)
2. What is the bound on the (asymptotic) rate of convergence?

Let S be a set of functions and let \mathcal{S} be an admissible structure. Consider now the case where the structure contains an infinite number of elements. Note that in this case in spite of the fact that any element S_k of the structure is characterized by a finite VC dimension h_k and a finite value B_k (finite value τ_k), the set of functions

$$S = \overline{\bigcup_{k=1}^{\infty} S_k}$$

can possess infinite VC dimension and/or infinite B_k (infinite τ_k).

We denote by $Q(z, \alpha_f^k)$, $k = 1, \dots$, the function which minimizes the empirical risk over the functions in the set S_k and denote by $Q(z, \alpha_0^k)$ the function which minimizes the expected risk over the functions in the set S_k ; we denote

also by $Q(z, \alpha_0)$ the function which minimizes the expected risk over the set of functions S .

In the following text, we prove the consistency of the SRM principle. However, first we show that there are rules for choosing the appropriate element S_n of the structure depending on the number of observations ℓ

$$n = n(\ell)$$

that provide risk convergence for chosen decision rules to the smallest possible risk.

For asymptotic results the refined bounds (6.8) are not very important. Therefore to simplify the exposition, consider instead of bound (6.8) the bound

$$R(\alpha) \leq R_{\text{emp}}(\alpha) + B_k \sqrt{\frac{h_k (\ln \frac{\ell}{k} + 1) - \ln \eta / 4}{\ell}} \quad (6.21)$$

that was obtained in Chapter 5, Section 5.3 for the pessimistic case.

Consider the a priori rule

$$n = n(\ell)$$

for choosing the number of element of the structure depending on the number of given examples.

Theorem 6.2. The *rule* $n = n(\ell)$ provides approximations $Q(z, a; '')$ for which the *sequence* of risks $R(\alpha_\ell^{n(\ell)})$ converges, as ℓ tends to infinity, to the smallest risk:

$$R(\alpha_0) = \inf_{\alpha \in A} \int Q(z, a) dP(z)$$

with asymptotic rate of convergence

$$V(\ell) = r_{n(\ell)} + \sqrt{\frac{D_{n(\ell)}^2 h_{n(\ell)} \ln \ell}{\ell}}, \quad (6.22)$$

where

$$r_{n(\ell)} = \int Q(z, \alpha_0^{n(\ell)}) dP(z) - \int Q(z, \alpha_0) dP(z), \quad (6.23)$$

(that is, the *equality*

$$P \left\{ \limsup_{\ell \rightarrow \infty} V^{-1}(\ell) \left| \int Q(z, \alpha_\ell^{n(\ell)}) dP(z) - \int Q(z, \alpha_0) dP(z) \right| < \infty \right\} = 1$$

holds true), if

$$\frac{D_{n(\ell)}^2 h_{n(\ell)} \ln \ell}{\ell} \xrightarrow[\ell \rightarrow \infty]{} 0, \quad n(\ell) \xrightarrow[\ell \rightarrow \infty]{} \infty, \quad (6.24)$$

where

- $D_n = B$, if one considers a structure with totally bounded functions $Q(z, a) \leq B$, in S_n and
- $D_n = \tau_n$ if one considers a structure with elements satisfying inequality (6.6).

The quantities

$$r_{n(\ell)} = \int Q(z, \alpha_0^{n(\ell)}) dP(z) - \inf_{\alpha \in \Lambda} \int Q(z, \alpha) dP(z)$$

describe the difference in risks between the smallest risk for a function from the element $S_{n(\ell)}$ of the structure \mathbf{S} and the smallest risk over the entire set of functions.

The next theorem is devoted to asymptotic properties of the structural risk minimization principle. It shows that if the SRM method uses a structure of elements that contains a totally bounded set of functions (see Section 1.1) then it is strongly universally consistent (that is, for any distribution function it provides convergence to the best possible solution with probability one).

To avoid choosing the minimum of functional (6.21) over the infinite number of elements of the structure, we introduce one additional constraint on the SRM method: We will choose the minimum from the first ℓ elements of the structure where ℓ is equal to the number of observations. Therefore we approximate the solution by function $Q(z, a)$, which among ℓ functions $Q(z, \alpha_\ell^k), k = 1, \dots, \ell$, minimizing empirical risk on corresponding elements $S_k, k = 1, \dots, \ell$, of the structure provide the smallest guaranteed (with probability $1 - 1/\ell$) risk:

$$R_{\text{emp}}^+(\alpha_\ell^+) = \min_{1 \leq k \leq \ell} \left[\frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha_\ell^k) + B_k \sqrt{\frac{h_k(\ln 2\ell/k + 1) + \ln 4\ell}{\ell}} \right].$$

The following theorem is an adaptation of the Lugosi–Zeger theorem for the set of not necessary indicator functions.

Theorem 6.3 (Lugosi, Zeger). *If the structure is such that $B_n^2 \leq n^{1-\delta}$, then for any distribution function the SRM method provides convergence to the best possible solution with probability one (i.e., the SRM method is universally strongly consistent).*

Moreover, if the optimal solution $Q(z, \alpha_0)$ belongs to some element S_* , of the structure $(Q(z, \alpha_0) = Q(z, \alpha_0^*))$ and $B_{n(\ell)}^2 \leq \mu(\ell) \leq \ell^{1-\delta}$, then using the SRM method one achieves the following asymptotic rate of convergence:

$$V(\ell) = O\left(\sqrt{\frac{\mu(\ell) \ln \ell}{\ell}}\right).$$

Remark. For the sets of indicator functions one chooses $B_n = 1$ for all n . In this case

$$\mu(\ell) = 1.$$

6.3.1 Proof of the Theorems

Proof of Theorem 6.2. We prove the theorem for the case $D_n = B_n$. The proof for the case $D_n = \tau_n$ is analogous.

Consider a structure with elements S_k containing totally bounded functions with the finite VC dimension. As shown in Section 5.3, for any element S_k with probability at least $1 - 2/\ell^2$ the additive bound

$$\Delta(\alpha_\ell^k) = R(\alpha_\ell^k) - R(\alpha_0^k) \leq B_k \left(\sqrt{\frac{-\ln \eta}{2\ell}} + \sqrt{\frac{h_k \left(\ln \frac{2\ell}{h_k} + 1 \right) + 2 \ln 2\ell}{\ell}} \right) \quad (6.25)$$

is valid. Then with probability $1 - 2/\ell^2$ the inequality

$$\begin{aligned} & R(\alpha_\ell^{n(\ell)}) - R(\alpha_0) \\ & \leq r_{n(\ell)} + B_{n(\ell)} \left(\sqrt{\frac{2 \ln \ell}{2\ell}} + \sqrt{\frac{h_{n(\ell)} \left(\ln \frac{2\ell}{h_{n(\ell)}} + 1 \right) + 2 \ln 2\ell}{\ell}} \right) \end{aligned} \quad (6.26)$$

holds, where

$$r_{n(\ell)} = R(\alpha_0^{n(\ell)}) - R(\alpha_0).$$

Since $S^* = \bigcup_k S_k$ everywhere dense in S , we have

$$\lim_{\ell \rightarrow \infty} r_{n(\ell)} = 0.$$

Therefore the condition

$$\lim_{\ell \rightarrow \infty} \frac{B_{n(\ell)}^2 h_{n(\ell)} \ln n(\ell)}{\ell} = 0$$

determines convergence to zero. Denote

$$V(\ell) = r_{n(\ell)} + B_{n(\ell)} \left(\sqrt{\frac{-\ln \eta}{2\ell}} + \sqrt{\frac{h_{n(\ell)} \left(\ln \frac{2\ell}{h_{n(\ell)}} + 1 \right) + 2 \ln 4\ell}{\ell}} \right)$$

Let us rewrite the assertion (6.26) in the form

$$P\{V^{-1}(\ell)(R(\alpha_\ell^{n(\ell)}) - R(\alpha_0)) > 1\} < \frac{2}{\ell^2}, \quad \ell > \ell_0.$$

Since

$$\sum_{\ell=1}^{\infty} P\{V^{-1}(\ell)(R(\alpha_\ell^{n(\ell)}) - R(\alpha_0)) > 1\} < \ell_0 + \sum_{\ell=\ell_0+1}^{\infty} \frac{2}{\ell^2} < \infty$$

according to the corollary from the Borel–Cantelli lemma (see Chapter 1, Section 1.11), one can assert that the inequality

$$\overline{\lim}_{\ell \rightarrow \infty} V^{-1}(\ell)(R(\alpha_\ell^{n(\ell)}) - R(\alpha_0)) \leq 1$$

is valid with probability one.

Proof of Theorem 6.3. Denote by α_ℓ^+ the parameter that minimizes guaranteed resk $R_{\text{emp}}^+(\alpha)$ using ℓ observations. Consider the decomposition

$$R(\alpha_\ell^+) - R(\alpha_0) = (R(\alpha_\ell^+) - R_{\text{emp}}^+(\alpha_\ell^+)) + (R_{\text{emp}}^+(\alpha_\ell^+) - R(\alpha_0))$$

For the first term of this decomposition we have

$$\begin{aligned} P\{R(\alpha_\ell^+) - R_{\text{emp}}^+(\alpha_\ell^+) > \varepsilon\} &< \sum_{k=1}^{\ell} P\{R(\alpha_\ell^k) - R_{\text{emp}}^+(\alpha_\ell^k) > \varepsilon\} \\ &= \sum_{k=1}^{\ell} P\left\{R(\alpha_\ell^k) - R_{\text{emp}}(\alpha_\ell^k) > \varepsilon + B_k \sqrt{\frac{h_k (\ln 2\ell/h_k + 1) + \ln 4\ell}{\ell}}\right\} \\ &\leq \sum_{k=1}^{\ell} 4 \left(\frac{2\ell e}{h_k}\right)^{h_k} \exp\left\{-\left(\frac{\varepsilon}{B_k} + \sqrt{\frac{h_k (\ln 2\ell/h_k + 1) + \ln 4\ell}{\ell}}\right)^2 \ell\right\} \\ &\leq \sum_{k=1}^{\ell} \frac{1}{\ell} \exp\left\{-\frac{\varepsilon^2 \ell}{B_k^2}\right\} \leq \exp\left\{-\frac{\varepsilon^2 \ell}{B_\ell^2}\right\} < \exp\left\{-\varepsilon^2 \ell^\delta\right\} \end{aligned}$$

where we take into account that $B_\ell^2 \leq \ell^{1-\delta}$. Using the Borel–Cantelli lemma we obtain that first summand of the decomposition converges almost surely to the nonpositive value.

Now consider the second term of the decomposition. Since S^* is dense everywhere in S , for every ε there exists an element S_s of the structure such that

$$R(\alpha_0^s) - R(\alpha_0) < \varepsilon.$$

Therefore we will prove that the second term in the decomposition does not exceed zero if we show that with probability one

$$\lim_{\ell \rightarrow \infty} \min_{1 \leq k \leq \ell} R_{\text{emp}}^+(\alpha_\ell^k) - R(\alpha_0^s) \leq 0$$

Note that for any ε there exists ℓ_0 such that for all $\ell > \ell_0$

$$B_s \sqrt{\frac{h_s(\ln 2\ell/h_s + 1) + 4 \ln \ell}{\ell}} \leq \frac{\varepsilon}{2}. \quad (6.27)$$

For $\ell > \ell_0$ we have

$$\begin{aligned} P \left\{ \min_{1 \leq k \leq \ell} R_{\text{emp}}^+(\alpha_\ell^k) - R(\alpha_0^s) > \varepsilon \right\} &\leq P \left\{ R_{\text{emp}}^+(\alpha_\ell^s) - R(\alpha_0^s) > \varepsilon \right\} \\ &= P \left\{ R_{\text{emp}}(\alpha_\ell^s) - R(\alpha_0^s) > \varepsilon - B_s \sqrt{\frac{h_s(\ln 2\ell/h_s + 1) + \ln 4\ell}{\ell}} \right\} \\ P \left\{ R_{\text{emp}}(\alpha_0^s) - R(\alpha_\ell^s) > \frac{\varepsilon}{2} \right\} &\leq P \left\{ \sup_{\alpha \in \Lambda} |R(\alpha) - R_{\text{emp}}(\alpha)| > \frac{\varepsilon}{2} \right\} \\ &< \left(\frac{2e\ell}{h_s} \right)^{h_s} \exp \left\{ -\frac{\varepsilon^2 \ell}{4B_s^2} \right\} \leq \left(\frac{2e\ell}{h_s} \right)^{h_s} \exp \left\{ -\frac{\varepsilon^2 \ell^\delta}{4} \right\}. \end{aligned}$$

Again applying the Borel–Cantelli lemma one concludes that second term of the decomposition converges almost surely to a nonpositive value. Since the sum of two terms is nonnegative, we obtain almost sure convergence $R(\alpha^+)$ to $R(\alpha_0)$. This proves the first part of the theorem.

To prove the second part, note that when the optimal solution belongs to one of the elements of the structure S_s the equality

$$R(\alpha_0^s) = R(\alpha_0)$$

holds true. Combining bounds for both terms, one obtains that for ℓ satisfying (6.27) the following inequalities are valid:

$$\begin{aligned} P \left\{ R(\alpha_\ell^+) - R(\alpha_0) > \varepsilon \right\} &\leq P \left\{ R(\alpha_\ell^+) - R_{\text{emp}}^+(\alpha_\ell^+) > \frac{\varepsilon}{2} \right\} + P \left\{ R(\alpha_\ell^+) - R(\alpha_0) > \frac{\varepsilon}{2} \right\} \\ &\leq \exp \left\{ -\frac{\varepsilon^2 \ell}{4\mu(\ell)} \right\} + \left(\frac{2e\ell}{h_s} \right)^{h_s} \exp \left\{ -\frac{\varepsilon^2 \ell}{16\mu(\ell)} \right\}. \end{aligned}$$

From this inequality we obtain the rate of convergence:

$$V(\ell) = O\left(\sqrt{\frac{\mu(\ell) \ln \ell}{\ell}}\right).$$

6.3.2 Discussions and Example

Thus, generally to estimate the asymptotic rate of convergence (6.22), one has to estimate two summands. The first summand

$$r_{n(\ell)} = \inf_{\alpha \in \Lambda_{n(\ell)}} \int Q(z, \alpha) dP(z) - \inf_{\alpha \in \Lambda} \int Q(z, \alpha) dP(z)$$

determines the rate of approximation—that is, the value of the deviation of the risk for the best approximation in S , from the smallest possible risk (the larger $n = n(\ell)$ the smaller is the deviation). The second summand

$$\sqrt{\frac{B_{n(\ell)}^2 h_{n(\ell)} \ln \ell}{\ell}}$$

determines the stochastic deviation of the risk obtained from the smallest risk in S :

$$\Delta_n(\ell) = \int Q(z, \alpha_\ell^{n(\ell)}) dP(z) - \inf_{\alpha \in \Lambda_{n(\ell)}} \int Q(z, \alpha) dP(z)$$

(the larger $n = n(\ell)$, the larger deviation $\Delta_n(\ell)$). Therefore the rate of convergence is determined by two contradictory requirements on the rule $n = n(\ell)$. For structures with a known bound on the rate of approximation, select the rule that assures the largest rate of convergence.

Section 6.5 discusses classical problems of function approximation. And shows that a good rate of approximation is possible only for special sets of functions (say for smooth functions). This fact is the main reason why in the general case the asymptotic rate of convergence for SRM can be slow.

However, in the particular case where the desired solution belongs to the element of the structure, the asymptotic rate of approximation is almost optimal

$$V(\ell) = O\left(\sqrt{\frac{\ln \ell}{\ell}}\right)$$

for the pattern recognition case, and it is arbitrarily close to this rate for estimating real-valued functions if one uses a structure with slowly increasing

bounds $B_{n(\ell)}^2 = \mu(\ell)$ (see Theorems 6.2 and 6.3):

$$V(\ell) = \left(\sqrt{\frac{\mu(\ell) \ln \ell}{\ell}} \right).$$

The following is an example of density estimation where the maximum likelihood method is not consistent but at the same time the method based on the SRM principle is consistent and has a high asymptotic rate of convergence.

Example. Let us consider a mixture of two normal laws

$$p(z; a, \sigma) = \frac{1}{2}N(a, \sigma) + \frac{1}{2}N(0, 1), \quad a \in (-\infty, \infty), \sigma \in (0, \infty), \quad (6.28)$$

where parameters a and σ of the first term of the law are unknown.

First we show that the maximum likelihood is not consistent for estimating these parameters from the sample

$$z_1, \dots, z_\ell, \dots$$

Indeed for any A and any ℓ , one can choose parameters a^* and σ^* such that

$$\sum_{i=1}^{\ell} \ln p(z_i; a^*, \sigma^*) > A.$$

This can be done, for example, by choosing $a = z_1$ and σ sufficiently small. For these parameters we have

$$\begin{aligned} & \sum_{i=1}^{\ell} \ln p(z_i; a^*, \sigma) \\ &= \ell \ln \frac{1}{2\sqrt{2\pi}} + \sum_{i=1}^{\ell} \ln \left(\exp \left\{ -\frac{z_i^2}{2} \right\} + \frac{1}{\sigma} \exp \left\{ -\frac{(z_i - z_1)^2}{2\sigma^2} \right\} \right) \\ &> \ell \ln \frac{1}{2\sqrt{2\pi}} + \ln \left(1 + \frac{1}{\sigma} \right) - \sum_{i=2}^{\ell} \frac{z_i^2}{2} \xrightarrow[\sigma \rightarrow 0]{} \infty. \end{aligned}$$

Thus the maximum likelihood does not yield the estimate of the desired parameters in the space (a, σ) .

Now let us use the structural risk minimization principle to modify the maximum likelihood method. We introduce the structure on the set of functions (6.28) and then minimize the guarantee risk over the elements of this structure.

To construct a structure, consider the sequence of positive values

$$b_1 > b_2 > \dots > b_n > \dots, \quad (6.29)$$

where

$$b_n = e^{-\sqrt{\mu(n)}}.$$

We define the following element S_k of the structure:

$$S_k = \{ \ln p(z, a, \sigma) : a \in (-\infty, \infty), \sigma \geq b_k \}.$$

These elements form sets of nested subsets. They satisfy all conditions to be an admissible structure: The **VC** dimension of any element of the structure does not exceed the finite **VC** dimension h of the set (6.28); all functions from the element S_k are bounded by the constant

$$B_n < \sqrt{\mu(n)};$$

the solution belongs to one of the elements of the structure.

Therefore for $n(\ell) = \ell$ we have the following rate:

$$\lim_{\ell \rightarrow \infty} \left(\frac{\mu(\ell) \ln \ell}{\ell} \right)^{-1/2} \int p(z, \alpha_0) \ln \frac{p(z, \alpha_\ell^+)}{p(z, \alpha_0)} dz < \infty,$$

which is the convergence (in the Kullback–Leibler metric) of the **SRM** estimates to the desired function with the asymptotic rate close to $\sqrt{\ln \ell / \ell}$ for slowly increasing function $\mu(\ell)$.

6.4 BOUNDS FOR THE REGRESSION ESTIMATION PROBLEM

In the previous section, we obtained bounds on the rate of convergence for the **SRM** principle. They have the order of magnitude

$$V(\ell) = O \left(r_{n(\ell)} + D_{n(\ell)} \sqrt{\frac{h_{n(\ell)} \ln \ell}{\ell}} \right), \quad (6.30)$$

where $D_s = B_s$ if the elements of the structure contain totally bounded functions and $D_u = \tau_n$ if the elements of structure contain unbounded functions.

This section considers an important special case; that is, we consider a model of regression estimation by series expansion using observations with additive noise. For this model we will obtain significantly better bounds. Under certain conditions it has the order of magnitude

$$V(\ell) = O \left(r_{n(\ell)} + \frac{h_{n(\ell)} \ln \ell}{\ell} \right).$$

6.4.1 The Model of Regression Estimation by Series Expansion

Let us specify the model. We consider a problem of estimating a regression function $f(x, \alpha_0) \in L_2(F)$, $x \in R^d$, where for any random vector x_i one has the measurement of the regression function $f(x, \alpha_0)$ with *additive noise* ξ :

$$\begin{aligned} y_i &= f(x_i, \alpha_0) + \xi_i, \\ E\xi &= 0, \quad E\xi^2 = \sigma^2, \quad E\xi_i \xi_j = 0 \quad \text{if } i \neq j. \end{aligned} \tag{6.31}$$

The problem is to estimate the regression function, using i.i.d. observations

$$(y_1, x_1), \dots, (y_\ell, x_\ell)$$

(here the x_i are random vectors, drawn according to the distribution function $F(x)$).

We define the structure \mathcal{S} using a set of complete orthonormal (with respect to probability measure $F(x)$) functions $\psi_k(x)$, $k = 1, 2, \dots$. The element S_k of this structure contains the functions of the form

$$f_k(x, \alpha) = \sum_{r=1}^k \alpha_r \psi_r(x).$$

Let the regression be described by the expansion in the series

$$f(x, \alpha_0) = \sum_{k=1}^{\infty} \alpha_k^0 \psi_k(x),$$

with an infinite number of terms.

We assume that the regression function has no singularities on this structure. This means that for all p the inequalities

$$\sup_x \left| \sum_{i=p+1}^{\infty} \alpha_i^0 \psi_i(x) \right| \leq c \tag{6.32}$$

hold true.

Let us denote

$$D_k = \left(\sup_{\alpha} \sup_{|\alpha|=1} \sum_{i=1}^k \alpha_i \psi_i(x) \right)^2. \tag{6.33}$$

As in Chapter 1 we determine the quality of approximation by the functional

$$R(\alpha) = \int (y - f(x, \alpha))^2 dF(x, y).$$

Let $f(x, \alpha, \cdot)$ be the function that minimizes the empirical risk

$$R_{\text{emp}}(\alpha) = \frac{1}{\ell} \sum_{i=1}^{\ell} (y_i - f(x_i, \alpha))^2$$

on the set of functions S_n and let

$$n = n(\ell)$$

be a rule for choosing the element of the structure S_n , depending on the number of examples. In this section we estimate the rate of convergence to zero of the quantities $R(\alpha_\ell^n) - R(\alpha_0)$. Taking (6.31) into account we obtain

$$R(\alpha_\ell^n) - R(\alpha_0) = \int (f(x, \alpha_\ell^n) - f(x, \alpha_0))^2 dF(x).$$

The following theorem holds true.

Theorem 6.4. *Let the model of regression estimation with additive noise satisfy the conditions (6.32) and (6.33). Then for any ℓ and for any rule $n = n(\ell)$ the SRM principle provides the bound*

$$P \left\{ V^{-1}(\ell) (R(\alpha_\ell^n) - R(\alpha_0)) \leq 1 \right\} \geq 1 - \frac{5}{\ln \ell}, \quad (6.34)$$

where

$$V(\ell) = r_n + \frac{n \ln \ell}{\ell \left(1 - \sqrt{D_n \mathcal{E}_n(\ell)} \right)_+^2} (\sigma^2 + c^2), \quad (6.35)$$

$$\mathcal{E}_n(\ell) = 4 \frac{n \left(\ln \frac{2\ell}{n} + 1 \right) + \ln \ell}{\ell}, \quad (6.36)$$

and

$$r_n = R(\alpha_0^n) - R(\alpha_0) = \int (f(x, \alpha_0^n) - f(x, \alpha_0))^2 dF(x).$$

Corollary. *If the rule $n = n(\ell)$ satisfies the condition*

$$\frac{D_{n(\ell)} n(\ell) \ln \ell}{\ell} \xrightarrow[\ell \rightarrow \infty]{} C < 1 \quad (6.37)$$

then the asymptotic rate of convergence has the order of magnitude

$$V(\ell) = O \left(r_{n(\ell)} + \frac{n(\ell) \ln \ell}{\ell} \right). \quad (6.38)$$

Note that the asymptotic bound (6.38) for this model is much better than the bound (6.30) for the general case.

Example. Let us estimate the regression function $f(x, \alpha_0)$ which is a periodic function and has $p > 0$ bounded derivatives, defined on the interval $(0, \pi)$. Let $F(x)$ be the uniform distribution on the interval $(0, \pi)$.

Consider the structure defined by the orthonormal series $\cos kx$, $k = 1, \dots$, where element S_n contains the functions

$$f_n(x, \alpha) = \sum_{k=1}^{\ell} \alpha_k \cos kx.$$

Since for a function that has $p > 0$ bounded derivatives the inequality

$$\sup_x \sum_{k=1}^{\infty} \alpha_k^0 \cos kx \leq \sum_{k=1}^{\infty} |\alpha_k^0| < \infty$$

holds, the nonsingularity condition (6.32) is satisfied.

For a given structure, one can easily find the bound

$$D_n = \left(\sup_x \sup_{|\alpha|=1} \sum_{k=1}^n \alpha_k \cos kx \right)^2 \leq \left(\sup_{|\alpha|=1} \sum_{k=1}^n \alpha_k \right)^2 \leq n.$$

Therefore according to Theorem 6.4 if a rule $n = n(\ell)$ for choosing the elements of the structure satisfies the condition

$$\frac{n^2(\ell) \ln \ell}{\ell} \xrightarrow{\ell \rightarrow \infty} 0 \quad (\text{condition (A)}),$$

then the following asymptotic rate of convergence for the obtained risk to the best possible holds true:

$$V(\ell) = r_n + \frac{n(\ell) \ln \ell}{\ell}$$

In the next section we describe the classical function approximation theorems according to which the rate of approximation by trigonometric series of risk functions that possess p derivatives has a risk

$$r_n = n^{-2p}(\ell).$$

Therefore

$$V(\ell) = n^{-2p}(\ell) + \frac{n(\ell) \ln \ell}{\ell}$$

One can easily verify that the rule

$$n = \left(\frac{\ell}{\ln \ell} \right)^{1/(2p+1)}$$

provides the best rate of convergence (this rule satisfies the condition (A) if $p \geq 1$). Using this rule, one obtains the asymptotic rate

$$V(\ell) = \left(\frac{\ln \ell}{\ell} \right)^{2p/(2p+1)};$$

that is,

$$\left(\frac{\ln \ell}{\ell} \right)^{-2p/(2p+1)} \int (f(x, \alpha_\ell^n(\ell)) - f(x, \alpha_0))^2 dx < \infty.$$

In the case where the regression function belongs to an element of the structure (in this case the regression function has an arbitrary number of bounded derivatives), by using a slowly increasing function $n = n(\ell)$ one achieves a rate of convergence that is close in order of magnitude to

$$V(\ell) = \left(\frac{\ln \ell}{\ell} \right).$$

6.4.2 Proof of Theorem 6.4

Denote by

$$f_n(x, \alpha_\ell) = \sum_{p=1}^{n(\ell)} \alpha_p \psi_p(x)$$

the function from the set of functions S_n that minimizes the empirical risk functional

$$R_{\text{emp}}(\alpha) = \frac{1}{\ell} \sum_i^{\ell} \left(y_i - \sum_{p=1}^{n(\ell)} \alpha_p \psi_p(x_i) \right)^2.$$

Let

$$f(x, \alpha_0) = \sum_{p=1}^{\infty} \alpha_p^0 \psi_p(x)$$

be the regression function. We estimate the rate of convergence in $L_2(F)$ of $f(x, \alpha_{n(\ell)})$ to the desired regression function

$$V(\ell) = \int \left(\sum_{p=1}^{\infty} \alpha_p^0 \psi_p(x) - \sum_{p=1}^{n(\ell)} \alpha_p^n \psi_p(x) \right)^2 dF(x)$$

Since the set of functions $\psi_p(x)$, $p = 1, 2, \dots$, is orthonormal with respect to the distribution $F(x)$, we obtain the following rate of approximation:

$$V(\ell) = \sum_p^{n(\ell)} \beta_p^2 + r_{n(\ell)}, \quad (6.39)$$

where we denote

$$\beta_p = \alpha_p^n - \alpha_p^0$$

and

$$r_{n(\ell)} = \sum_{p=n(\ell)+1}^{\infty} (\alpha_p^0)^2$$

To bound the sum (6.39) we have to bound the first term:

$$T_1(\ell) = \sum_p^{n(\ell)} \beta_p^2.$$

To do so we define a vector $\beta = (\beta_1, \dots, \beta_n)$ corresponding to α_p which minimizes the empirical risk

$$\begin{aligned} R_{\text{emp}}(\beta) &= \frac{1}{\ell} \sum_{i=1}^{\ell} \left(y_i - \sum_{p=1}^{n(\ell)} \alpha_p \psi_p(x_i) \right)^2 \\ &= \frac{1}{\ell} \sum_{i=1}^{\ell} \bar{y}_i^2 - 2 \sum_{p=1}^{n(\ell)} \beta_p G_p + \sum_{p,q=1}^{n(\ell)} \beta_p \beta_q \frac{1}{\ell} \sum_{i=1}^{\ell} \psi_p(x_i) \psi_q(x_i), \end{aligned}$$

where we denote

$$\begin{aligned} G_p &= \frac{1}{\ell} \sum_{i=1}^{\ell} \bar{y}_i \psi_p(x_i), \\ \bar{y}_i &= \xi_i + \sum_{p=n+1}^{\infty} \alpha_0^p \psi_p(x_i). \end{aligned} \quad (6.40)$$

Denote by K the covariance matrix with elements

$$K_{p,q} = \frac{1}{\ell} \sum_{i=1}^{\ell} \psi_p(x_i) \psi_q(x_i)$$

and denote by $G = (G_1, \dots, G_n)^T$ the n -dimensional vector of coordinates (6.40). In these notations the vector $\beta(n(\ell))$ minimizing the empirical risk is

$$\beta_{n(\ell)} = K^{-1}G.$$

Therefore the bound

$$T_1(\ell) = |\beta_{n(\ell)}|^2 = |K^{-1}G|^2 \leq |K^{-1}|^2|G|^2 \quad (6.41)$$

holds true. Let us bound the norm of the matrix K^{-1} and the norm of vector G from above. The norm of matrix K equals μ_{\max}^n , the largest eigenvalue of K , and the norm of the matrix K^{-1} :

$$|K^{-1}| = \frac{1}{\mu_{\min}^n},$$

where μ_{\min}^n is the smallest eigenvalue of the $n \times n$ matrix K . Therefore to bound K^{-1} we have to bound μ_{\min}^n from below.

Consider the function

$$\Phi_n(x, \alpha) = \left(\sum_{p=1}^n \alpha_p \psi_p(x) \right)^2, \quad (6.42)$$

which we shall examine in the domain

$$\sum_{p=1}^n \alpha_p^2 = 1. \quad (6.43)$$

Recall that we have defined the bound D_n such that

$$\sup_x \Phi_n(x, \alpha) \leq D_n$$

in the domain (6.43). Now consider the expression

$$\frac{1}{\ell} \sum_{i=1}^{\ell} \Phi_n(x_i, \alpha) = \frac{1}{\ell} \sum_{i=1}^{\ell} \left(\sum_{p=1}^n \alpha_p \psi_p(x_i) \right)^2.$$

Observe that

$$E \Phi_n(x, \alpha) = \sum_{p=1}^n \alpha_p^2, \quad (6.44)$$

$$\frac{1}{\ell} \sum_{i=1}^{\ell} \Phi_n(x_i, \alpha) = \sum_{p=1}^n \alpha_p \alpha_q K_{p,q}, \quad (6.45)$$

where $K_{p,q}$, $p, q = 1, \dots, n$, are the elements of covariance matrix K described above. Using a rotation transformation, we arrive at a new orthogonal system of functions $\psi_1^*(x), \dots, \psi_n^*(x)$ such that

$$E\Phi_n(x, \alpha^*) = \sum_{p=1}^n (\alpha_p^*)^2, \quad (6.46)$$

$$\frac{1}{\ell} \sum_{i=1}^{\ell} \Phi_n(x_i, \alpha^*) = \sum_{p,q} \mu_p(\alpha_p^*)^2, \quad (6.47)$$

where μ_1, \dots, μ_n are eigenvalues of the matrix K .

To bound the eigenvalues use the results of the Theorem 5.3 according to which with probability at least $1 - 4/\ell$ simultaneously for all functions $\Phi_n(x, \alpha)$ in the domain (6.43) the following inequality holds true:

$$E\Phi_n(x, \alpha^*) \leq \frac{1}{\ell} \sum_{i=1}^{\ell} \Phi_n(x_i, \alpha^*) + \frac{D_n \mathcal{E}_n}{2} \left(1 + \sqrt{1 + \frac{4 \sum_{i=1}^{\ell} \Phi_n(x_i, \alpha^*)}{\ell D_n \mathcal{E}_n(\ell)}} \right), \quad (6.48)$$

where for our structure with $h_n = n$ we have

$$\mathcal{E}_n(\ell) = 4 \frac{n \left(\ln \frac{2\ell}{n} + 1 \right) + \ln \ell}{\ell}$$

Taking into account (6.46) and (6.47) we can rewrite this inequality for domain (6.43):

$$\sum_{p=1}^n (1 - \mu_p)(\alpha_p^*)^2 \leq \frac{D_n \mathcal{E}_n}{2} \left(1 + \sqrt{1 + \frac{4 \sum_{p=1}^n (\alpha_p^*)^2 \mu_p}{D_n \mathcal{E}_n(\ell)}} \right).$$

This inequality is true with probability $1 - 4/\ell$ simultaneously for all α_p^* in domain (6.43). Therefore with probability $1 - 4/\ell$ the bound is valid for the specific vector $\alpha = (0, 0, \dots, 1, 0, \dots, 0)$ where the one corresponds to the smallest eigenvalue. For this vector we obtain the following inequality:

$$\mu_{\min}^n \geq 1 - \frac{D_n \mathcal{E}_n(\ell)}{2} \left(1 + \sqrt{1 + \frac{4 \mu_{\min}^n}{D_n \mathcal{E}_n(\ell)}} \right)$$

where

$$\mu_{\min}^n = \min_{1 \leq p \leq n} \mu_p^n.$$

Solving this inequality with respect to μ_{\min}^n , one obtains that with probability at least $1 - 1/\ell$ the inequality

$$\mu_{\min}^n > \left(1 - \sqrt{D_n \mathcal{E}_n(\ell)} \right)_+ \quad (6.49)$$

holds true, where we define $(u)_+ = \max(u, 0)$.

Therefore with probability $1 - 1/\ell$ the bounds

$$|K^{-1}|^2 \leq \frac{1}{(1 - \sqrt{D_n \mathcal{E}_n(\ell)})_+^2} \quad (6.50)$$

hold true.

To bound $|G|^2$ note that

$$|G|^2 = \sum_{p=1}^n G_p^2 = \sum_{p=1}^n \frac{1}{\ell^2} \left(\sum_{i=1}^{\ell} \bar{y}_i \psi_p(x_i) \right)^2.$$

Let us compute the expectation

$$\begin{aligned} E|G|^2 &= \sum_{p=1}^n EG_p^2 \\ &= \sum_{p=1}^n \frac{1}{\ell^2} E \left(\sum_{i=1}^{\ell} \psi_p(x_i) \left(\xi_i + \sum_{j=p+1}^{\infty} \alpha_j^0 \psi_j(x_i) \right) \right)^2 \leq n \frac{\sigma^2 + c^2}{\ell}. \end{aligned} \quad (6.51)$$

To derive the inequality (6.51) we use the condition (6.32) that the regression function has no singular structure.

To bound the random value $|G|$ we utilize Chebyshev inequality for the first moments

$$P \{ \xi > \epsilon \} < \frac{E\xi}{\epsilon},$$

where we use

$$\epsilon = \frac{n \ln \ell}{\ell} (\sigma^2 + c^2).$$

We obtain

$$P \left\{ |G|^2 > \frac{n(\sigma^2 + c^2) \ln \ell}{\ell} \right\} < \frac{1}{\ln \ell}.$$

Thus with probability at least $1 - 1/\ln \ell$

$$|G|^2 \leq \frac{n \ln \ell}{\ell} (\sigma^2 + c^2) \quad (6.52)$$

holds true. Substituting (6.50) and (6.52) in (6.41) we obtain that with probability at least $1 - 2/\ln \ell$ the first term in the sum (6.31) is bounded as follows:

$$T(n) \leq \frac{n \ln \ell}{\ell(1 - \sqrt{D_n \mathcal{E}_n(\ell)})_+^2} (\sigma^2 + c^2).$$

Therefore we proved that

$$P \left\{ V^{-1}(\ell) (R(\alpha_\ell^n) - R(\alpha_0)) \leq 1 \right\} \geq 1 - \frac{2}{\ln \ell},$$

where

$$V(\ell) = r_n + \frac{n \ln \ell (\sigma^2 + c^2)}{\ell(1 - \sqrt{D_n \mathcal{E}_n(\ell)})_+^2}$$

The theorem has been proved.

6.5 THE PROBLEM OF APPROXIMATING FUNCTIONS

In the previous sections, we obtained the asymptotic rate of convergence for the SRM principle. We showed that in the general case the asymptotic rate of convergence has order of magnitude

$$V(\ell) = r_{n(\ell)} + D_{n(\ell)} \sqrt{\frac{h_{n(\ell)} \ln \ell}{\ell}}, \quad (6.53)$$

where $D_n = B_n$ if the elements of the structure contain totally bounded functions and $D_n = \tau_n$ if elements of the structure contain an unbounded set of functions.

For the problem of regression estimation, with quadratic loss function we obtained the bound which (under some conditions) has a better order of magnitude:

$$V(\ell) = r_{n(\ell)} + \frac{h_{n(\ell)} \ln \ell}{\ell} \quad (6.54)$$

To use these results, however, one needs to estimate the first term in (6.53)

and (6.54) that describes the rate of convergence of the risks r_n attained at the best function of the elements of the structure to the smallest possible risk for entire set of function. This section is devoted to estimating this rate of convergence (rate of approximation).

Note that in the general case we have

$$\begin{aligned} r_n &= \int Q(z, \alpha_0^n) dF(z) - \int Q(z, \alpha_0) dF(z) \\ &\leq \inf_{\alpha \in \Lambda_n} \int |Q(z, \alpha_0) - Q(z, \alpha)| dF(z). \end{aligned}$$

Let us denote the right-hand side of this inequality by r_n^* :

$$r_n^* = \inf_{\alpha \in \Lambda_n} \int |Q(z, \alpha_0) - Q(z, \alpha)| dF(z).$$

The quantities r_n^* describe a rate of approximation in the (weak) metric $L_1(F)$ of the desired function $Q(z, \alpha_0)$ by the best functions of the elements of the structure S .

For the case of measurements with additive noise (6.31) and quadratic loss functions

$$Q(z, \alpha) = (y - f(x, \alpha))^2, \quad \alpha \in \Lambda$$

the rate of convergence of the risks

$$\begin{aligned} r_n &= \int Q(z, \alpha_0^n) dF(z) - \int Q(z, \alpha_0) dF(z) \\ &= \int (f(x, \alpha_0) - f(x, \alpha_0^n))^2 dF(x) = r_n^* \end{aligned}$$

coincides with the square of the rate of function approximation in $L_2(F)$ metric.

Therefore to estimate the rate of risk convergence in $L_2(F)$ metric it is sufficient to estimate the rate of function approximation for the corresponding structure.

Estimating the rate of function approximation (in different metrics, not necessarily in weak ones[†]) constitutes the problem of approximation theory. This theory was started more than 100 years ago when Weierstrass had discovered that on the finite interval every continuous function admits approximation to any accuracy by algebraic polynomials. This posed the question: How fast do polynomial approximations converge to the desired function with increasing degree of polynomials?

[†]Note that the rate of function approximation in weak metrics is not worse than the rate of approximation in the strong metric C.

The approximation theory addresses the following problem. Let Φ be a set of functions belonging to a normed space of functions. Consider the structure

$$\mathcal{M}_1 \subset \mathcal{M}_2 \subset \cdots \subset \mathcal{M}_n, \dots \quad (6.55)$$

imposed on this set with the following property: The elements of the structure $\{\mathcal{M}\}_{k=1}^{\infty}$ are such that $\bigcup_{k=1}^{\infty} \mathcal{M}_k$ is dense in Φ .

The problem is for a given set of functions Φ and for a given structure to find a bound

$$\rho(f, \mathcal{M}_k) = \inf_{f^* \in \mathcal{M}_k} \|f - f^*\| \leq r_n^*,$$

which is valid for any function f of the set Φ .

This, however, is a very general setting of the problem. Approximation theory considers some special sets of functions Φ and some special structures $\{\mathcal{M}\}_{k=1}^{\infty}$ for which it estimates the approximation rate.

In the following sections, we will state (without proofs) some classical theorems of approximation theory, then we will formulate the theorems estimating the rate of convergence for the structures used in the learning models, and lastly we will demonstrate some connections between the rate of approximation and the VC dimension of the set of approximated functions.

6.5.1 Three Theorems of Classical Approximation Theory

This section describes three theorems of constructive approximation theory—that is, the theory that not only gives a bound of the approximation rate, but also provides the methods that for any given function f how to find in the subset \mathcal{M}_k the best approximation f_k^* .

Let us consider the classical problem of approximation of the periodic functions $f(x)$, $x \in \mathbb{R}'$, by the Fourier series. We will approximate the function $f(x)$ by the trigonometric series

$$\Phi_k(x) = \frac{a_0}{2} + \sum_{j=1}^k (a_j \cos jx + b_j \sin jx), \quad (6.56)$$

where

$$a_0 = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) dx,$$

$$a_j = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos jx dx, \quad (6.57)$$

$$b_j = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin jx dx. \quad (6.58)$$

An important role in Fourier analysis is played by the Dirichlet formula

$$D_N(x) = \frac{1}{2} + \sum_{j=1}^N \cos jx = \frac{\sin\left(\frac{(N+1)x}{2}\right)}{\sin\left(\frac{x}{2}\right)} \quad (6.59)$$

The right-hand side of the expression (6.59) is called the *Dirichlet kernel*. Using kernel $D_k(x)$ one can rewrite the approximation (6.56) in the form

$$\Phi_N(x) = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x - \tau) D_N(\tau) d\tau. \quad (6.60)$$

However, the approximation (6.60) does not guarantee convergence to any point of continuous target function.

Therefore along with the Dirichlet kernel $D_k(x)$ one considers the so-called *Fejer kernel*

$$\mathcal{F}_N(x) = \frac{1}{N} \sum_{k=0}^{N-1} D_k(x) = \frac{\sin^2 \frac{Nx}{2}}{2N \sin^2 \frac{x}{2}} \quad (6.61)$$

This kernel defines Fejer approximations

$$F_N(x) = \int_{-\pi}^{\pi} f(x - \tau) \mathcal{F}_N(\tau) d\tau \quad (6.62)$$

in the Fourier expansion. Note that the Fejer approximation gives the expression

$$F_N(x) = \frac{a_0^*}{2} + \sum_{j=1}^N a_j^* \cos jx + \sum_{j=1}^N b_j^* \sin jx,$$

where coefficients a_i^* and b_i^* are regularized Fourier coefficients ((6.57) and (6.58))

$$a_i^* = \left(1 - \frac{i}{N}\right) a_i, \quad b_i^* = \left(1 - \frac{i}{N}\right) b_i.$$

In 1904, Fejer proved that any continuous function, periodic on finite interval, can be approximated by the Fejer approximation (6.62) with any degree of accuracy as N increases. However, on the basis of Fejer approximation one cannot determine the rate of approximation.

In 1911, Jackson gave a construction that guarantees the best asymptotic rate of approximation (in the metric C) of $r \geq 0$ times continuously differentiable periodic functions.

To formulate Jackson's theorem we need to define the following concept.

Definition. We call the quantity

$$\omega(\delta, f(x)) = \sup_{|h| \leq \delta} \sup_x \{ |f(x + h) - f(h)| \}$$

the *modulus of continuity of the function* $f(x)$.

Now let us define the Jackson kernel

$$J_{N,r}(x) = \lambda_{N,r} \left(\frac{\sin \frac{\left[\frac{N}{r} \right] x}{2}}{\sin \frac{x}{2}} \right)^{2r}, \quad r = 2, 3, \dots, \quad (6.63)$$

where coefficient $\lambda_{N,r}$ is chosen to normalize the kernel

$$\int_{-\pi}^{\pi} J_{N,r}(x) dx = 1.$$

It follows from (6.63) that the kernel $J_{N,r}(x)$ is an even, nonnegative trigonometric polynomial of degree $\leq N$.

Now let us describe an approximating function $f_N(x)$ from \mathcal{M}_N . We will distinguish between two cases.

In the first case, we have no information about the smoothness properties of the desired function. In this case, we construct the approximation

$$f_N(x) = \pi^{-1} \int J_{N,2}(x) f(x + \tau) d\tau$$

using the Jackson kernel $J_{N,2}(x)$ with $r = 2$.

In the second case we have information that the desired function has no less than $r > 1$ derivatives. In this case we will construct the approximation

$$f_N(x) = \pi^{-1} \int J_{N,r}(x) f(x + \tau) d\tau$$

using the Jackson kernel $J_{N,r}(x)$ with parameter r .

Theorem 6.5 (Jackson). *Let $f(x)$ be an r times continuously differentiable periodic function. Then the inequality*

$$\rho_C(f(x), f_{N,r}(x)) \leq A^{r+2} N^{-r} \omega(N^{-1}, f^{(r)}(x)) \quad (6.64)$$

holds true, where $A < \pi\sqrt{3}/2$ is a universal constant.

The converse theorem is valid as well.

Theorem 6.6 (Bernstein (1912) and Vallee-Poussin (1919)). Let a continuous function $f(x)$ satisfy the inequality

$$\rho_C(f, \mathcal{M}_N) \leq C(f)N^{-(r+\delta)}, \quad (6.65)$$

where r is some integer, $0 < \delta < 1$ and $\rho_C(f, \mathcal{M}_N)$ is distance between the function f and the closest function from the set of trigonometric polynomials of degree N in C metric,

$$\rho_C(f(x), \mathcal{M}_N) = \inf_{f^* \in \mathcal{M}_N} \sup_x |f(x) - f^*(x)|.$$

Then $f(x)$ is r times differentiable and its r th derivative satisfies a Lipschitz condition of order δ

$$|f^{(r)}(x) - f^{(r)}(x')| \leq A|x - x'|^\delta.$$

These two theorems show that the rate of approximation by Fourier sums depends solely on smoothness properties of the target function; the smoother the target function, the higher the rate of approximation by trigonometric sums.

The same result remains true if one considers a structure with elements \mathcal{M}_N containing algebraic polynomials of degree N .

Theorem 6.7. Let $f(x)$ be an $r \geq 0$ times continuously differentiable function on $[a,b]$ and let function $f^{(r)}(x)$ satisfy the Lipschitz condition of order δ :

$$|f^{(r)}(x) - f^{(r)}(x')| \leq A|x - x'|^\delta \quad \text{for } x, x' \in [a, b].$$

Then the inequality

$$\rho_C(f(x), \mathcal{M}_N) \leq C(f)N^{-(r+\delta)} \quad (6.66)$$

holds true, where the constant $C(f)$ depends on the function $f(x)$.

6.5.2 Curse of Dimensionality in Approximation Theory

Now we have to generalize the results obtained for the one-dimensional case to the multidimensional case.

Let Φ^* be a set of functions defined on the d -dimensional cube $[0,1]^d$ and let functions from this set have bounded (in the uniform norm) partial derivatives of order s and satisfy the (d -dimensional) Lipschitz condition of order $0 < \delta < 1$. Consider the following structure: Element \mathcal{M}_n is the set of

polynomials of degree n in each of the d variables that is linear in parameter space of dimension $N_n = d^n$ (here N_n is the number of parameters).

Theorem 6.8. For any function $f(x)$ of the set Φ^* the following inequality holds:

$$P_C(f, \mathcal{M}_n) \leq C(f)N_n^{-(s+\delta)/d} \quad (6.67)$$

where constant $C(f)$ depends on function f

From (6.67) we find that the asymptotic rate of convergence drastically decreases with increasing number of parameters when the characteristic of smoothness (number of bounded derivatives) remains fixed.

Therefore according to approximation theory one can guarantee good approximation of a high-dimensional function only if the desired function is extremely smooth.

6.5.3 Problem of Approximation in Learning Theory

In the learning theory we have to estimate the rate of approximation even for more difficult cases.

We have to estimate the rate of approximation for the cases when:

1. Φ is a set of high-dimensional functions.
2. The elements \mathcal{M}_n of the structure are not necessarily linear manifolds.
They can be any sets of functions with finite VC dimension.

Furthermore, we are interested in the cases where the rate of approximation is rather high (otherwise one cannot hope to find a good approximation of the desired function using a restricted number of observations). We will call the rate of approximation high if it has a bound $O(1/\sqrt{n})$, where n is an index of an element of the structure.

Therefore in the learning theory we face a problem: to describe cases for which the high rate of approximation is possible. This means to describe different sets of smooth functions and structures for these sets that provide the bound $O(1/\sqrt{n})$.

Below we consider a new concept of smoothness. Let $\{f(x)\}$ be a set of functions and let $\{\bar{f}(\omega)\}$ be a set of their Fourier transforms.

We will characterize the smoothness of the function $f(x)$ by the value b such that

$$\int |\omega|^b \bar{f}(\omega) d\omega = C_b(f) < \infty, \quad b \geq 0. \quad (6.68)$$

In terms of this concept the following theorems hold true

Theorem 6.9 (Jones, 1992). Let the set of functions $f(x)$ satisfy (6.68) with $b = 0$. Consider the structure with elements \mathcal{M}_n containing the functions

$$f(x) = \sum_{i=1}^n c_i \sin((x, w_i) + v_i), \quad (6.69)$$

where c_i and v_i are arbitrary values and w_i are arbitrary vectors. Then the rate of approximation of the desired function by the best function of the elements (6.69) in L_2 metric is bounded by $O(1/\sqrt{n})$.

Theorem 6.10 (Barron, 1993). Let the set of functions $f(x)$ satisfy (6.68) with $b = 1$. Consider the structure with elements \mathcal{M}_n containing the functions

$$f(x) = \sum_{i=1}^n c_i \delta((x, w_i) + v_i), \quad (6.70)$$

where c_i and v_i are arbitrary values and w_i is an arbitrary vector, $\delta = \delta(u)$ is a sigmoid function: (a monotonic increasing function such that

$$\lim_{u \rightarrow -\infty} \delta(u) = -1, \quad \lim_{u \rightarrow \infty} \delta(u) = 1.$$

Then the rate of approximation of the desired function by the best functions of the elements (6.70) in L_2 metric is bounded by $O(1/\sqrt{n})$.

Theorem 6.11 (Breiman, 1993). Let the set of functions $f(x)$ satisfy (6.68) with $b = 2$. Consider the structure with elements \mathcal{M}_n containing the functions

$$f(x) = \sum_{i=1}^n c_i \sin|x \cdot w_i + v_i|_+ + x \cdot a + b, \quad |u|_+ = \max(0, u), \quad (6.71)$$

where c_i and v_i and b are arbitrary values and w_i and a are arbitrary vectors. Then the rate of approximation of the desired function by the best function of the elements (6.71) in L_2 metric is bounded by $O(1/\sqrt{n})$.

In spite of the fact that in these theorems the concept of smoothness differs from the number of bounded derivatives, one can observe the similar phenomenon as in the classical case: To keep a high rate of convergence in a space with increasing dimensionality, one has to increase the smoothness property of the function. Using concept (6.68), one attains it automatically. Girosi and Anzellotti (1993) showed that a set of functions satisfying (6.68) with $b = 1$ and $b = 2$ can be rewritten, respectively, as

$$f(x) = \frac{1}{|x|^{n-1}} * \lambda(x), \quad f(x) = \frac{1}{|x|^{n-2}} * \lambda(x),$$

where A is any function whose Fourier transform is integrable, and $*$ stands for the convolution operator. In this form it becomes more apparent that functions satisfying (6.68) become more and more constrained as the dimensionality increases due to more rapid fall-off of the terms $1/|x|^{n-1}$ and $1/|x|^{n-2}$.

Therefore if the desired function is not very smooth, one cannot guarantee high asymptotic rate of convergence of the constructed approximations to the desired function.

6.5.4 The VC Dimension in Approximation Theory

In this section we will describe a special class of sets of functions for which the rate of approximation is high and the bounds depend on the VC dimension of some set of functions.

Consider set of functions which is defined by the functions $\lambda(t)$ belonging to L_1 and some fixed kernel $K(x, t)$:

$$f(x) = \int K(x, t)\lambda(t) dt, \quad (6.72)$$

where $x, t \in R^n$, and the kernel $K(x, t)$ satisfies the condition

$$|K(x, t)| \leq \tau.$$

In this representation by using different kernel functions one transforms functions from L_1 into different sets of (smooth) functions.

Let us rewrite (6.72) in the form

$$f(x) = \int K^*(x, t)p(t) dt,$$

where

$$K^*(x, t) = |\lambda| \text{sign}(\lambda(t)) K(x, t),$$

$$|\lambda| = \int |\lambda(t)| dt$$

and

$$p(t) = \frac{|\lambda(t)|}{|\lambda|}$$

is some density function. Therefore (using the results obtained in Chapter 5, Section 5.5) one can assert that if we sample ℓ points t_1, \dots, t_ℓ from $p(t)$, then with probability $1 - \eta$ we have

$$\sup_x \left| f(x) - \frac{1}{\ell} \sum_{i=1}^{\ell} K^*(x, t_i) \right| \leq 2|\lambda|\tau\sqrt{\mathcal{E}(\ell)},$$

where

$$\mathcal{E} = \frac{h \left(\ln \frac{2\ell}{h} + 1 \right) - \ln \eta/4}{\ell} + \frac{1}{\ell},$$

and h is VC dimension of the set of functions $K^*(x, t)$ (here t describes the vector of variables and x describes the vector of parameters).

Since for any positive η there exist ℓ points t_1^*, \dots, t_ℓ^* that satisfy this inequality, the inequality

$$\sup_x \left| f(x) - \frac{1}{\ell} \sum_{i=1}^{\ell} K^*(x, t_i^*) \right| \leq 2|\lambda|\tau \sqrt{\mathcal{E}^*(\ell)}$$

where

$$\mathcal{E}^*(\ell) = \frac{h \left(\ln \frac{2\ell}{h} + 1 \right) + \ln 4}{\ell} + \frac{1}{\ell}$$

holds true with probability one.

Thus we have proved the following theorem.

Theorem 6.12 (Girosi). *Let $f(x)$ be a set of functions, defined by representation (6.72). Let the kernel $K(x, t)$ be considered as a parametric set of functions (with respect to the variables t) that has the VC dimension h .*

Consider the structure S with elements S_N containing the functions

$$f(x) = |\lambda| \sum_{i=1}^N \frac{c_i}{N} K(x, t_i),$$

defined by the parameters

$$|\lambda| \in \mathbb{R}^1, \quad t_i \in \mathbb{R}^d, \quad c_i \in \{-1, 1\} \quad i = 1, \dots, N.$$

Then the rate of approximation in metric C of any function from this set by the elements of the structure is

$$r_N^* \leq 2|\lambda|\tau \sqrt{\frac{h \left(\ln \frac{2N}{h} + 1 \right) + \ln 4}{N}}.$$

This theorem describes the way of constructing the sets of functions and appropriate structures for which the rate of approximation in C metric is high.

As an example of application of this idea, let us make the following choice of the kernel function:

$$K(x, t) = G_m(x - t),$$

where $G_m(u)$ is the Bessel potential function of the order m which has the following integral representation (Stein, 1970, p. 132):

$$G_m(u) = \frac{(2\pi)^{-m}}{\Gamma(m/2)} \int_0^\infty e^{-\sigma/4\pi} \sigma^{(m-d-2)/2} e^{-(\pi/\sigma)|u|^2} d\sigma$$

To be bounded at the origin, this kernel must satisfy the condition

$$m > d.$$

It is known that the space of functions

$$f = G_m * \lambda \quad (6.73)$$

with $\Lambda \in L_1$ forms the so-called Sobolev–Liouville space L_1^m (space of functions with absolutely integrable m th derivative).

Let us rewrite the representation (6.73) in the explicit form

$$f(x) = \frac{(2\pi)^{-m}}{\Gamma(m/2)} \int e^{-(\pi/\sigma)|x-t|^2} \Lambda(t, \sigma) dt d\sigma,$$

where we denote

$$\Lambda(t, \sigma) = e^{-(\sigma/4\pi)} \sigma^{(m-n-2)/2} \lambda(t).$$

To apply the results of Theorem 6.12 to this specific kernel, estimate the VC dimension of the set of functions

$$\phi(x, t, \sigma) = e^{-(\pi/\sigma)|x-t|^2}, \quad x, t \in R^d, \beta \in R^1.$$

Note that the VC dimension of this set of real functions is equivalent to the VC dimension of the set of indicator functions

$$\psi(x, t, \sigma) = \theta\{|x-t|^2 - \beta\sigma\},$$

which is equal to $d + 1$.

Thus, we have proved the following theorem.

Theorem 6.13 (Girosi). Let f be an element of space L_1^m , with $m > d$.

Consider the *structure* S with elements S_N containing the functions

$$f_N(x) = |\Lambda| \frac{(2\pi)^{-m}}{\Gamma(m/2)} \sum_{i=1}^N \frac{c_i}{N} e^{-(\pi/\sigma_i)|x-t_i|^2},$$

defined by parameters

$$|\Lambda| \in R^1, \quad t_i \in R^d, \quad c_i \in \{-1, 1\}.$$

Then the rate of function approximation in the metric C of any function from this set by the elements of the structure is

$$r_N^* = 2|\Lambda|\tau \sqrt{\frac{(d+1) \left(\ln \frac{2N}{d+1} + 1 \right) + \ln 4}{N}}$$

Note how clearly this theorem confirms the statement that a high asymptotic rate of approximation can be achieved only for smooth functions (the functions considered in Theorem 6.13 belong to Sobolev space with $m > d$).

Note also that according to Theorem 6.12 one can construct smooth functions by using convolution with kernel of very general type. To get high asymptotic rate of approximation the only constraint is that the kernel should be a bounded function which can be described as a family of functions possessing finite VC dimension.

6.6 PROBLEM OF LOCAL RISK MINIMIZATION

In the last section, we made a rather pessimistic conclusion: The high *asymptotic* rate of approximation can be achieved only for very smooth functions.

Note, however, that this assertion was made for the asymptotic rate of approximation. It does not exclude that nonasymptotic behavior of approximation can be much better.

Recall that we have already seen this situation. The main bound for the method of minimizing empirical risk over a set of totally bounded functions with finite VC dimension is the following: With probability $1 - \eta$ the inequality

$$R(\alpha_\ell) \leq R_{\text{emp}}(\alpha_\ell) + \frac{B\mathcal{E}(\ell)}{2} \left(1 + \sqrt{1 + \frac{4R_{\text{emp}}(\alpha_\ell)}{B\mathcal{E}(\ell)}} \right) \quad (6.74)$$

holds true, where

$$\mathcal{E}(\ell) = 4 \frac{h \left(\ln \frac{\ell}{h} + 1 \right) - \ln \eta / 4}{\ell}$$

In the most typical real-life cases when both $\ell/h > 1$ and the amount $R_{\text{emp}}(\alpha_\ell)/B\mathcal{E}(\ell)$ are small, the rate of convergence behaves as $\mathcal{E}(\ell)$. However, if $R(\alpha_0) \neq 0$, then asymptotically for large ℓ/h the rate of convergence behaves as $\sqrt{\mathcal{E}(\ell)}$. (When $R(\alpha_0) = R_{\text{emp}}(\alpha_\ell) = 0$ we have a special case where the rate equals $\mathcal{E}(\ell)$.)

Thus, for estimating the function from a restricted amount of observations, one has to find a structure that provides small values $R_{\text{emp}}(\alpha_\ell^n)$ when ℓ/h is small. It is clear, however, that in order to find this structure, one has to possess some prior information about the problem at hand.

To decrease the influence of choosing a poor structure, we will consider a new statement of the problem: the problem of local estimation of a function, that is the estimation of function in a vicinity of the point of interest.

However, before moving into the theory of local function estimation, let us clarify why local function estimation can be better than global.

Suppose we would like to estimate the function plotted on Fig 6.3a. Suppose we are given the structure where element S_k is a set of polynomials of order k . As shown in Fig 6.3a, to approximate this function well on the interval $[0,1]$ we have to use a polynomial of high degree (to describe well a flat part of the curve). Therefore we need a polynomial of high degree m to obtain an appropriate level of approximation.

Now let us consider two problems: estimating the desired function on the interval $[0, 1/2]$ and estimating this function on the interval $[1/2, 1]$.

For this example a reasonable level of accuracy of approximation of the desired function can be achieved by approximating the function on the interval $[0, 1/2]$ by polynomial of degree 0 and approximating the desired function on the interval $[1/2, 1]$ by a polynomial of degree 1 (see Fig 6.3b). In other words, better accuracy can be achieved by approximating the function locally.

In general, it is possible to obtain an additional gain if one can make a "smart" partition of the interval $[0, 1]$ into (two or more) subintervals.

The problem arises: *How do we partition a (multidimensional) input space into subspaces to obtain a good solution for problem of local function estimation?*

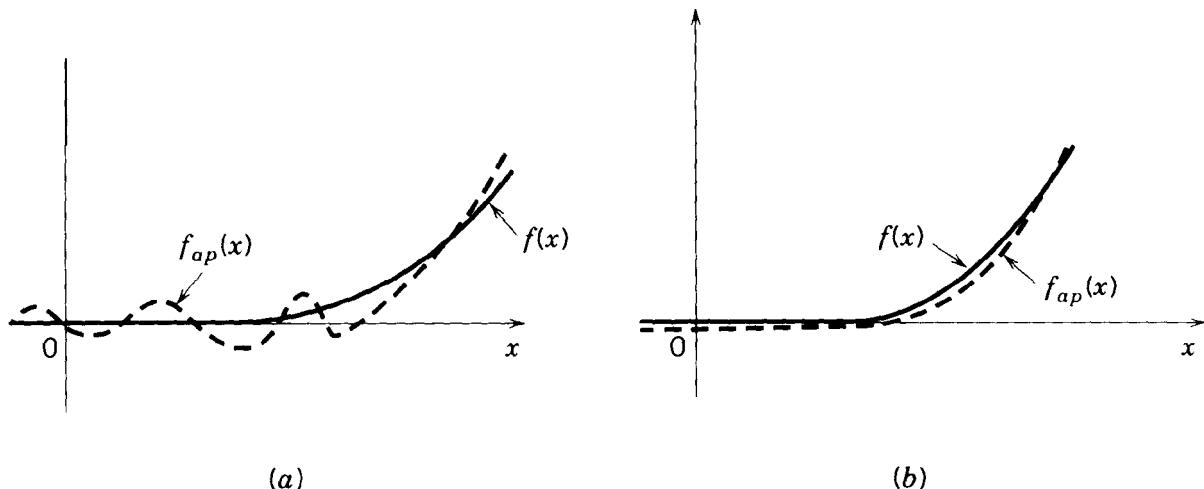


FIGURE 6.3. (a) To approximate function well on interval $(0,1)$, one needs a polynomial of high degree. (b) To approximate the same function well on the two semi-intervals, one needs a low degree polynomials.

To consider such a problem we introduce the model of the local risk minimization.

6.6.1 Local Risk Minimization Model

In all our previous considerations we used a loss function defined by some variable $z = (y, x)$. Now to introduce the specific structure of loss functions we consider two variables y and x . Consider a nonnegative function $K(x, x_0; \beta)$ that embodies the concept of vicinity. This function depends on a point x_0 and on a "locality" parameter $\beta \in (0, \infty)$ and satisfies two conditions:

$$\begin{aligned} 0 &\leq K(x, x_0; \beta) \leq 1, \\ K(x_0, x_0; \beta) &= 1. \end{aligned} \quad (6.75)$$

For example, both the "hard threshold" vicinity function (Fig. 6.4a)

$$K_1(x, x_0; \beta) = \begin{cases} 1 & \text{if } ||x - x_0|| < \frac{\beta}{2}, \\ 0 & \text{otherwise} \end{cases} \quad (6.76)$$

and the "soft threshold" vicinity function (Fig. 6.4b)

$$K_2(x, x_0; \beta) = \exp \left\{ -\frac{(x - x_0)^2}{\beta^2} \right\} \quad (6.77)$$

meet these conditions. Let us define the value

$$\mathcal{K}(x_0, \beta) = \int K(x, x_0; \beta) dF(x). \quad (6.78)$$

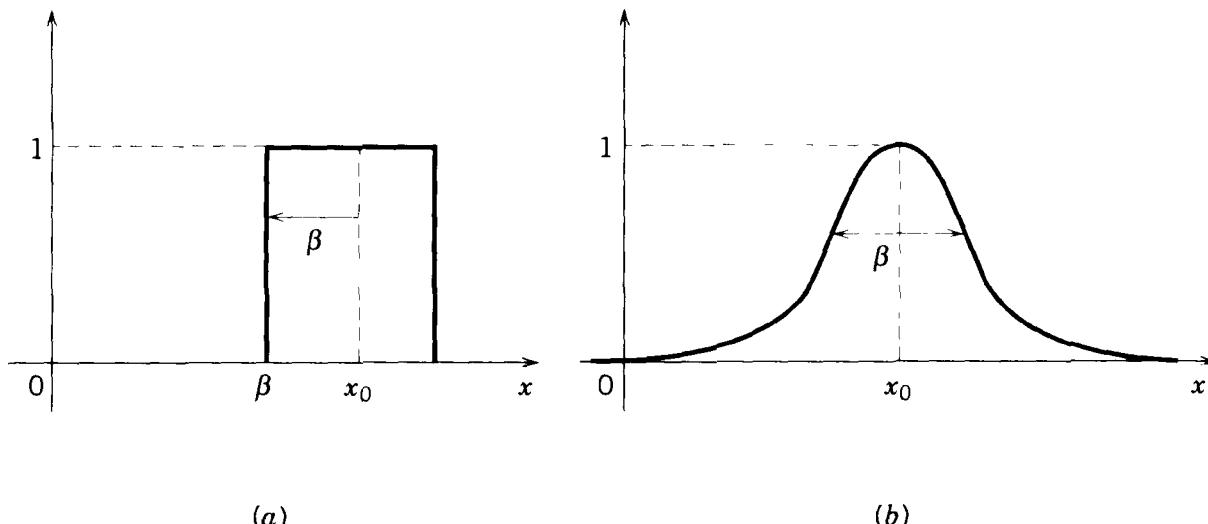


FIGURE 6.4. Examples of vicinity functions. (a) Hard-threshold vicinity function. (b) Soft-threshold vicinity function.

For the set of functions $f(x, a)$, $a \in A$, let us consider the loss functions $Q(z, a) = Q(y, f(x, a))$, $a \in A$. Our goal is to minimize the local **risk** functional

$$R(\alpha, \beta; x_0) = \int Q(y, f(x, \alpha)) \frac{K(x, x_0; \beta)}{K(x_0; \beta)} dF(x, y) \quad (6.79)$$

over both the set of functions $f(x, a)$, $a \in A$, and different vicinity functions at the point x_0 (defined by parameter β) in a situation where the probability measure $F(x, y)$ is unknown, but we are given examples

$$(x_1, y_1); \dots; (x_\ell, y_\ell).$$

Note that the problem of local risk minimization on the basis of empirical data is a generalization of the problem of global risk minimization. (In the last problem we have to minimize the functional (6.79), where $K(x, x_0; \beta) = 1$.)

Let us apply the statement of local risk minimization problem to our specific problems: the pattern recognition problem and the problem of regression estimation.

In the simplest setting of the problem of pattern recognition both the set of functions $f(x, a)$, $a \in A$, and the set of loss functions $Q(y, f(x, a))$ are sets of indicator functions

$$Q(y, f(x, \alpha)) = \begin{cases} 0 & \text{if } y = f(x, \alpha), \\ 1 & \text{if } y \neq f(x, \alpha). \end{cases} \quad (6.80)$$

Minimizing the local risk functional (6.79) with a hard threshold vicinity function for this problem means to find both the vicinity of point x_0 (parameter β^* for inequality $\|x - x_0\| \leq \frac{\beta^*}{2}$) and the function $f(x, a')$ which minimize the probability of error in the region $\|x - x_0\| \leq \frac{\beta^*}{2}$.

For the problem of regression estimation we consider the following loss functions:

$$Q(y, f(x, \alpha)) = (y - f(x, \alpha))^2, \quad (6.81)$$

where now $f(x, a)$, $a \in A$, is a set of real-valued functions and y is a real value.

Remark. It looks reasonable to choose the point x_0 of interest as a center of admissible vicinities and then use the obtained function for estimating the value of the desired function at the point x_0 .

Note that in this statement for different points of interest x_0 one has to estimate both the approximating function and the value of vicinity.

We will distinguish between two different models of estimating the function in the vicinity of a given point:

1. The case where for any given point of interest x_0 we choose both the value of the vicinity $K(x, \beta_\ell)$ and the approximating function $f(x, \alpha_\ell)$. We call this case a local approximation of the desired function.
2. The case where for different points of interest we use the same value of the vicinity but different approximating functions. We call this case a *semilocal* approximation of the desired function.

In the following examples we show that well-known classical methods such as the method of K-nearest neighbors in pattern recognition and the Watson–Nadaraya method (method of moving average) in the regression estimation problem are the simplest semilocal methods for solving the problem of minimizing the local risk functional (6.79), which uses the empirical risk minimization principle.

Example 1. Consider the pattern recognition problem. Our goal is using the empirical data to minimize the functional (6.79) with loss function (6.80), where:

1. $K(x, x_0, \beta)$ is the hard threshold vicinity function (6.76) and
2. $f(x, a), a \in A$, is the simplest set of indicator functions, namely the *set* of constant functions.

Suppose we are given the value of the vicinity parameter $\beta = \beta^*$, which we will use for every possible point of interest (we have a fixed parameter of vicinity).

To minimize this functional, let us minimize the empirical risk functional

$$R_{\text{emp}}(\alpha, x_0, \beta^*) = \frac{1}{\ell K(x_0, \beta^*)} \sum_{i=1}^{\ell} Q(y_i, f(x_i, \alpha)) K(x_i, x_0, \beta^*)$$

over the set of constant functions $f(x, a) = c$. Note that in the set of indicator functions there exist only two constant functions, namely the function $f(x, \alpha_1) \equiv 1$ and the function $f(x, \alpha_2) \equiv 0$.

To minimize the empirical risk for this case means to check how many errors these two functions make on the part of the training set that falls into the vicinity of the point of interest and then assign the point of interest to those constant functions that make less errors.

Thus we obtained the method of type K-nearest neighbors.[†] This method is semilocal because the value of the vicinity is fixed a priori.

[†]The difference is only that here the vicinity of the point of interest x_0 is determined by the radius of the sphere with center in this point, and therefore the number of elements of the training set in the vicinity of the point of interest is not fixed. In the classical K-nearest neighbor method the vicinity of the point of interest x_0 is measured by the radius such that the corresponding sphere includes a given numbers of elements of the training set and therefore the radius is not fixed (but the number of points in the sphere is fixed).

Example 2. Now let us consider the problem of regression estimation. Our goal is to minimize the functional (6.79), with loss function (6.81) where:

1. $K(x, x_0, \beta)$ is a soft-threshold vicinity function and
2. $f(x, a)$, $a \in A$, is the simplest set of real-valued functions, namely the *set of constant functions*.

As in Example 1 we are given a priori the value of the vicinity parameter $\beta = \beta^*$.

The subset of constant functions of the set of real-valued function contains an infinite number of elements $f(x, a) = a$, $a \in (-\infty, \infty)$.

To minimize this functional, let us minimize the empirical risk functional

$$R(\alpha, x_0, \beta^*) = \frac{1}{\ell K(x_0, \beta^*)} \sum_{i=1}^{\ell} (y_i - \alpha)^2 K(x_i, x_0, \beta^*)$$

over parameters α . The minimum of this functional is achieved for the constant function

$$\alpha^* = \sum_{i=1}^{\ell} \frac{K(x_i, x_0, \beta^*)}{\sum_{i=1}^{\ell} K(x_i, x_0, \beta^*)} y_i. \quad (6.82)$$

Using this constant function we evaluate the value of the desired function at the point x_0 . Therefore Eq. (6.82) gives a function $\phi(x)$ for estimating the value of the desired function in any given point of interest x :

$$\phi(x) = \sum_{i=1}^{\ell} \frac{K(x_i, x, \beta^*)}{\sum_{i=1}^{\ell} K(x_i, x, \beta^*)} y_i. \quad (6.83)$$

Expression (6.83) is called the Watson–Nadaraya estimator or the moving average estimator. The same as the K-nearest neighbor estimator, this estimator is semilocal.

6.6.2 Bounds for the Local Risk Minimization Estimator

This section presents three theorems concerning the bounds for the local risk functional. Using these bounds, one can utilize the structural risk minimization principle for minimizing the local risk functional on the basis of empirical data. Note that the principle of minimizing the empirical risk

$$R_{\text{emp}}(\alpha, x_0, \beta^*) = \frac{1}{\ell K(x_0, \beta^*)} \sum_{i=1}^{\ell} Q(y, f(x)) K(x_i, x_0, \beta^*) \quad (6.84)$$

over two parameters α and β^* gives a bad solution.

Indeed, for a hard-threshold vicinity function and for the set of constant functions the minimum of empirical risk is equal to zero if β is such that the vicinity of the point of interest includes only one element of training set. However, this solution does not guarantee generalization.

We derive bounds on risk that are valid simultaneously for all sets of functions and all values of the vicinity parameter.

We start with the case where $Q(y, f(x, a)), a \in A$, is a set of indicator functions which has VC dimension h_1 .

Theorem 6.14. *Let the set of indicator functions $Q(y, f(x, a)), a \in A$, have the VC dimension h_1 and let the set of nonnegative real-valued functions $K(x, x_0, \beta), \beta \in (0, \infty)$, have the VC dimension h_2 . Then with probability $1 - 2\eta$ simultaneously for all $a \in A$ and all $\beta \in (0, \infty)$ the inequality*

$$R(\alpha, \beta, x_0) \leq \frac{2R_{\text{emp}}(\alpha, \beta, x_0) + \mathcal{E}_{h_1}(\ell) + \mathcal{E}_{h_2}(\ell)}{2(\mathcal{K}_{\text{emp}}(x_0, \beta) - \sqrt{\mathcal{E}_{h_2}(\ell)})_+} \left(1 + \sqrt{1 + \frac{4R_{\text{emp}}(\alpha, \beta, x_0)}{\mathcal{E}_{h_1}(\ell) + \mathcal{E}_{h_2}(\ell)}} \right) \quad (6.85)$$

holds true, where

$$R_{\text{emp}}(\alpha, \beta, x_0) = \frac{1}{\ell} \sum_{i=1}^{\ell} Q(y_i, f(x_i, \alpha)) K(x_i, x_0, \beta), \quad (6.86)$$

$$\mathcal{E}_{h_i}(\ell) = 4 \frac{h_i \left(\ln \frac{2\ell}{h_i} + 1 \right) - \ln \eta / 4}{\ell}, \quad i = 1, 2, \quad (6.87)$$

$$\mathcal{K}_{\text{emp}}(x_0, \beta) = \frac{1}{\ell} \sum_{i=1}^{\ell} K(x_i, x_0, \beta). \quad (6.88)$$

Remark. The bound (6.85) uses the VC dimension of two sets of functions: VC dimension h_1 of the set of indicator functions $Q(x, f(x, a)), a \in A$, and VC dimension h_2 of the set of real-valued functions $K(x, x_0, \beta), \beta \in (0, \infty)$. The numerator of the bound of the risk depends on the sum $\mathcal{E}_{h_1}(1) + \mathcal{E}_{h_2}(\ell)$ (the smaller the sum, the smaller the numerator), and the denominator of the bound depends only on one parameter $\mathcal{E}_{h_2}(\ell)$ (the smaller $\mathcal{E}_{h_2}(\ell)$, the larger the denominator). Therefore, it seems advisable to define a set of vicinity functions $K(x, x_0, \beta), \beta \in (0, \infty)$, with small VC dimension—for example, *monotonic radial functions*. The VC dimension of the set of radial functions

$$K(x, x_0, \beta) = K_\beta(\|x - x_0\|), \quad \beta \in (0, \infty),$$

where $K_\beta(r)$ are monotonically nonincreasing functions of r , is equal to 1. For this set of functions we have a bound

$$\begin{aligned} & R(\alpha, \beta, x_0) \\ & \leq \frac{2R_{\text{emp}}(\alpha, \beta, x_0) + \mathcal{E}_{h_1}(\ell) + \mathcal{E}_1(\ell)}{2 \left(\mathcal{K}_{\text{emp}}(x_0, \beta) - \sqrt{\mathcal{E}_1(\ell)} \right)_+} \left(1 + \sqrt{1 + \frac{4R_{\text{emp}}(\alpha, \beta, x_0)}{\mathcal{E}_{h_1}(\ell) + \mathcal{E}_1(\ell)}} \right), \end{aligned} \quad (6.89)$$

where $R_{\text{emp}}(\alpha, \beta)$, $\mathcal{E}_h(\ell)$, and $\mathcal{K}_{\text{emp}}(x_0, \beta)$ are determined by Eqs. (6.86), (6.87), and (6.88).

The next theorem is a generalization of Theorem 6.14 for the case where $Q(y, f(x, \alpha))$, $\alpha \in A$, is a set of real-valued totally bounded nonnegative functions $0 \leq Q(y, f(x, \alpha)) \leq B$.

Theorem 6.15. *Let $0 \leq Q(y, f(x, \alpha)) \leq B$, $\alpha \in A$, be a set of totally bounded nonnegative functions, let the set of functions $Q(y, f(x, \alpha)) \times K(x, x_0, \beta)$, $\alpha \in A$, $\beta \in (0, \infty)$, have the VC dimension h^* , and let the set of functions $K(x, x_0, \beta)$, $\beta \in (0, \infty)$, have the VC dimension h_2 . Then with probability $1 - 2\eta$ simultaneously for all $\alpha \in A$ and all $\beta \in (0, co)$ the inequality*

$$\begin{aligned} & R(\alpha, \beta, x_0) \\ & \leq \frac{2R_{\text{emp}}(\alpha, \beta, x_0) + B\mathcal{E}_{h^*}(\ell)}{2 \left(\mathcal{K}_{\text{emp}}(x_0, \beta) - \sqrt{B\mathcal{E}_{h_2}(\ell)} \right)_+} \left(1 + \sqrt{1 + \frac{4R_{\text{emp}}(\alpha, \beta, x_0)}{B\mathcal{E}_{h^*}(\ell)}} \right) \end{aligned} \quad (6.90)$$

is valid where $R_{\text{emp}}(\alpha, \beta)$, $\mathcal{E}_h(\ell)$ and $\mathcal{K}_{\text{emp}}(x_0, \beta)$ are determined by Eqs. (6.86) (6.87), and (6.88).

Remark. In contrast to the set of real-valued functions $Q(y, f(x, \alpha)) \times K(x, x_0, \beta)$, $\alpha \in A$, $\beta \in (0, co)$, considered in Theorem 6.14 where any function was a product of an indicator function and a real-valued function, here we consider a set of real-valued functions which are products of two real-valued functions.

For the specific set of real-valued functions considered in Theorem 6.14, we obtained a bound which depends on the VC dimensions of the two set of each factors separately.

For the functions considered in Theorem 6.15, this assertion is not true. For example, let $\{\phi^*(z)\}$ be the set of all monotonically increasing functions in $z \in \mathbb{R}^1$ and $\phi_*(z)$ be a set of all monotonically decreasing functions in $z \in \mathbb{R}^1$. Although the VC dimension of both sets is 1, the VC dimension of the product of these sets $\phi^*(z)\phi_*(z)$ is infinite. Therefore in the bound of Theorem 6.15 we have to use the VC dimension of the product of the two sets.

Lastly, consider the bounds on the local risk functional for the case of the set of unbounded loss functions $Q(y, f(x, a))$, $a \in A$.

In Chapter 5 we considered the bound on the risk for an unbounded set of nonnegative functions, satisfying the condition

$$\sup_{\alpha \in \Lambda} \frac{\sqrt{\int Q^p(z, \alpha) dF(z)}}{\int Q(z, \alpha) dF(z)} \leq \tau, \quad p > 1.$$

Here we consider a stronger restriction. To describe it we note that the expression

$$dF_\beta(z) = \frac{K(x, x_0, \beta)}{\mathcal{K}(x_0, \beta)} dF(z), \quad \beta \in (0, \infty)$$

defines a family of measures that depends on parameter β . Below we assume that inequality

$$\sup_{\beta} \sup_{\alpha \in \Lambda} \frac{\sqrt{\int Q^2(z, \alpha) dF_\beta(z)}}{\int Q(z, \alpha) dF_\beta(z)} \leq \tau, \quad (6.91)$$

holds true,[†] where the supremum is taken both over set of functions $Q(z, a)$, $a \in A$, and the family of measures $F_\beta(z)$.

Theorem 6.16. *Let the set of nonnegative functions $Q(y, f(x, \alpha))$ $x K(x, x_0, \beta)$, $a \in A$, $\beta \in (0, \infty)$, have VC dimension h^* and satisfy inequality (6.91). Let the family of vicinity functions $K(x, x_0, \beta)$, $\beta \in (0, \infty)$, have VC dimension h_2 . Then with probability $1 - 2\eta$ simultaneously for all $a \in A$ and all $\beta \in (0, \infty)$ the inequality*

$$R(\alpha, \beta, x_0) \leq \frac{R_{\text{emp}}(\alpha, \beta, x_0)}{\left(\mathcal{K}_{\text{emp}}(x_0, \beta) - \sqrt{\mathcal{E}_{h_2}(\ell)} \right)_+ \left(1 - \tau \sqrt{\frac{\mathcal{E}_{h^*}(\ell)(1 - \ln \mathcal{E}_{h^*}(\ell))}{(\mathcal{K}_{\text{emp}}(x_0, \beta) - \sqrt{\mathcal{E}_{h_2}(\ell)})_+}} \right)_+} \quad (6.92)$$

holds true, where $\mathcal{E}_h(\ell)$ is determined by the expression (6.87).

6.6.3 Proofs of the Theorems

To prove Theorem 6.14 note that:

1. The set of real-valued functions $Q(y, f(x, a))$ $x K(x, x_0, \beta)$, $a \in A$, is totally bounded by the constant 1.

[†]We consider case $p = 2$ only to simplify notations in the formulas for the bounds of the local risk. One can easily obtain the bounds for any $p > 1$.

2. The VC dimension of the set of real-valued functions $Q(y, f(x, \alpha))_x K(x, x_0, \beta)$ is equal to the VC dimension of the following set of indicator functions

$$\Phi(x, y, \alpha, \beta) = \theta(Q(y, f(x, \alpha)) \times K(x, x_0, \beta) - \gamma), \quad (6.93)$$

where $0 \leq y \leq 1$. Recall that we consider the case where $Q(y, f(x, \alpha))$, $\alpha \in A$, is a set of indicator functions. It is easy to check that for this set of functions the equality

$$\theta(Q(y, f(x, \alpha)) \times K(x, x_0, \beta) - \gamma) = Q(y, f(x, \alpha))\theta(K(x, x_0, \beta) - \gamma) \quad (6.94)$$

holds true. Let the VC dimension of the set of indicator functions $Q(y, f(x, \alpha))$, $\alpha \in A$, be h_1 and let VC dimension of the set of indicator functions $\theta[K(x, x_0, \beta) - \gamma]$, $\beta \in (0, \infty)$, $y \in [0, 1]$, be h_2 .

Note that the growth function for the set of indicator functions which are the products of indicator functions from two different sets does not exceed the product of the two growth functions:

$$G^{\Lambda, R^1}(t) \leq G^\Lambda(\ell)G^{R^1}(\ell) \leq \left(\frac{e\ell}{h_1}\right)^{h_1} \left(\frac{e\ell}{h_2}\right)^{h_2}, \quad (6.95)$$

where $G^{\Lambda, R^1}(\ell)$ is the growth function of the set (6.94), $G^\Lambda(\ell)$ is the growth function of the set $Q(z, \alpha)$, $\alpha \in A$, and $G^{R^1}(\ell)$ is growth function for the set $\theta[K(x, x_0, \beta) - \gamma]$, $\beta \in (0, \infty)$, $\gamma \in [0, 1]$.

Let us denote

$$R^*(\alpha, \beta, x_0) = \int Q(y, f(x, \alpha))K(x, x_0, \beta) dF(x, y). \quad (6.96)$$

Using results obtained in Chapter 4, Section 4.6 we derive that with probability at least $1 - \eta$ the inequality

$$R^*(\alpha, \beta, x_0) \leq R_{\text{emp}}(\alpha, \beta, x_0) + \frac{\mathcal{E}_{h_1}(\ell) + \mathcal{E}_{h_2}(\ell)}{2} \left(1 + \sqrt{1 + \frac{4R_{\text{emp}}(\alpha, \beta, x_0)}{\mathcal{E}_{h_1}(\ell) + \mathcal{E}_{h_2}(\ell)}}\right)$$

holds true simultaneously for all $\alpha \in A$ and all $\beta \in (0, \infty)$, where $\mathcal{E}_{h_i}(\ell)$ is given by (6.87).

Dividing both sides of the inequality by $K(x_0, \beta)$ we obtain

$$R(\alpha, \beta, x_0) \leq \frac{2R_{\text{emp}}(\alpha, \beta, x_0) + \mathcal{E}_{h_1}(\ell) + \mathcal{E}_{h_2}(\ell))}{2K(x_0, \beta)} \left(1 + \sqrt{1 + \frac{4R_{\text{emp}}(\alpha, \beta, x_0)}{\mathcal{E}_{h_1}(\ell) + \mathcal{E}_{h_2}(\ell)}}\right). \quad (6.97)$$

To prove the theorem, it is sufficient to obtain a lower bound on the quantity $\mathcal{K}(x_0, \beta)$. Since the value of the functions $K(x, x_0, \beta)$ does not exceed 1 and this set of functions has VC dimension h_2 , we obtain

$$P\left\{\sup_{\beta}|\mathcal{K}(x_0, \beta) - \mathcal{K}_{\text{emp}}(x_0, \beta)| > \varepsilon\right\} < 4 \exp\left\{\left(\frac{h_2 \left(\ln \frac{2\ell}{h_2} + 1\right)}{\ell} \varepsilon^2\right) \ell\right\}$$

From this inequality we derive that with probability $1 - \eta$ simultaneously for all β the inequality

$$\mathcal{K}(x_0, \beta) \geq \mathcal{K}_{\text{emp}}(x_0, \beta) - \sqrt{\mathcal{E}_{h_2}(\ell)} \quad (6.98)$$

holds true.

Using this inequality and inequality (6.97), we obtain the assertion of the theorem.

The proof of Theorem 6.15 is analogous.

To prove Theorem 6.16 note that according to Theorem 5.4 (for $p = 2$) the following inequality is valid:

$$\begin{aligned} P\left\{\sup_{\alpha, \beta} \frac{R(\alpha, \beta, x_0) - R_{\text{emp}}(x_0, \alpha, \beta)}{\sqrt{\int Q^2(y, f(x, \alpha)) K^2(x, x_0, \beta) dF(x, y)}} > \varepsilon \sqrt{1 - \frac{\ln \varepsilon}{2}}\right\} \\ < \exp\left\{\left(\frac{h^* \left(\ln \frac{2\ell}{h^*} + 1\right)}{\ell} - \frac{\varepsilon^2}{4}\right) \ell\right\}. \end{aligned} \quad (6.99)$$

Since $0 \leq K(x, x_0, \beta) \leq 1$ and inequality (6.91) is valid, we have

$$\begin{aligned} & \sqrt{\int Q^2(y, f(x, \alpha)) K^2(x, x_0, \beta) dF(x, y)} \\ & \leq \sqrt{\int Q^2(y, f(x, \alpha)) K(x, x_0, \beta) dF(x, y)} \\ & \leq \sqrt{\mathcal{K}(x_0, \beta) \int Q^2(y, f(x, \alpha)) dF_{\beta}(x, y)} \\ & \leq \tau \sqrt{\mathcal{K}(x_0, \beta)} \int Q(y, f(x, \alpha)) dF_{\beta}(x, y) \\ & = \tau \frac{\int Q(y, f(x, \alpha)) K(x, x_0, \beta) dF(x, y)}{\sqrt{\mathcal{K}(x_0, \beta)}}. \end{aligned} \quad (6.100)$$

Using (6.99) and (6.100), we reinforce inequality

$$\begin{aligned} P \left\{ \sup_{\alpha, \beta} \frac{R(\alpha, \beta, x_0) - R_{\text{emp}}(x_0, \alpha, \beta)}{R(\alpha, \beta, x_0)} \sqrt{\mathcal{K}(x_0, \beta)} > \tau \varepsilon \sqrt{1 - \frac{\ln \varepsilon}{2}} \right\} \\ < 4 \exp \left\{ \left(\frac{h^*(\ln \frac{2\ell}{h^*} + 1)}{\ell} - \frac{\varepsilon^2}{4} \right) \ell \right\}. \end{aligned}$$

Rewriting this in the equivalent form with probability $1 - \eta$ simultaneously for all α and β we obtain the inequality

$$R(\alpha, \beta, x_0) \leq \frac{R_{\text{emp}}(\alpha, \beta, x_0)}{\mathcal{K}(x_0, \beta) \left(1 - \tau \sqrt{\frac{\mathcal{E}_{h^*}(\ell)(1 - \ln \mathcal{E}_{h^*}(\ell))}{\mathcal{K}(x_0, \beta)}} \right)_+}. \quad (6.101)$$

To prove the theorem it remains to substitute the lower bound (6.98) for $\mathcal{K}(x_0, \beta)$ in (6.102).

6.6.4 Structural Risk Minimization Principle for Local Function Estimation

Using the obtained bound, one can apply the SRM principle to the local function estimation problem using the bounds provided by Theorems 6.14, 6.15, and 6.16. Let us start with the pattern recognition problem.

Consider a nested structure on the set of indicator functions $Q(y, f(x, a))$, $a \in A$.

$$S_1 \subset S_2 \subset \dots \subset S_N.$$

Let the VC dimension of each subset S_k of functions be $h_1(k)$, such that

$$h_1(1) \leq h_1(2) \leq \dots \leq h_1(N).$$

According to Theorem 6.14 with probability $1 - 2q$ simultaneously for all functions $Q(y, f(x, a))$, $a \in \Lambda_k$, from S_k and all vicinity functions the bound (6.85) is valid.

Since this bound is valid for all functions of the set S_k , it is valid for the function $Q(y, f(x, \alpha_\ell^k))K(x, x_0, \beta_\ell^k)$, which minimizes the right-hand side of inequality simultaneously over both parameters $a \in \Lambda_k$ and $\beta \in (0, 1)$ as well. The choice of the function $Q(y, f(x, \alpha_\ell^k))$ and the vicinity function $K(x, x_0, \beta_\ell^k)$ guarantees with probability $1 - 2\eta$ the smallest local risk for the function of element S_k .

Therefore consider each element S_k , and then choose a pair composed of the best function $Q(y, f(x, a:))$ and the best vicinity function $K(x, x_0, \beta_\ell^k)$.

Now it remains to choose the element and the corresponding vicinity function which provide the smallest bound.

The scheme of structural risk minimization for regression is identical. As in the pattern recognition case, we consider the structure \mathbf{S} with nested elements S_k that contain sets of totally bounded functions $Q(y, f(x, \alpha)) \times K(x, x_0, \beta)$ with the common constant B_k (or the sets of unbounded functions satisfying (6.91) with the common constant τ_k).

To choose the best pair we minimize the right-hand side of the bound (6.90) given by Theorem 6.15 (or the right-hand side of the bound (6.92) given by Theorem 6.16 if the elements of the structure contains unbounded functions).

Note that using local risk minimization methods, one probably does not need rich sets of approximating functions (recall that the classical semi-local methods are based on using a set of constant functions). For local estimation functions in the one-dimensional case, it is probably enough to consider elements S_k , $k = 0, 1, 2, 3$, containing the polynomials of degree 0, 1, 2, 3.

APPENDIX TO CHAPTER 6: ESTIMATING FUNCTIONS ON THE BASIS OF INDIRECT MEASUREMENTS

This appendix applies the SRM principle to the problem of estimating the function on the basis of results of indirect measurements. Although this problem belongs to the so-called stochastic ill-posed problems whose theory we consider in the next chapter, the particular setting of this problem considered here has some singularities. These singularities allow us to solve the problem of indirect measurements using the structural risk minimization principle.

A6.1 PROBLEMS OF ESTIMATING THE RESULTS OF INDIRECT MEASUREMENTS

Let it be required to estimate the function $f(t, \alpha_0) = f(t)$ in the set of functions $f(t, a)$, $a \in A$ (here $f(t)$ belongs to $f(t, a)$), in a situation where it is impossible to measure directly the values of the function $f(t)$, but one can measure the values of an another function $F(x)$, $a \leq x \leq b$, related to the desired one by means of the operator equation

$$Af(t) = F(x). \quad (\text{A6.1})$$

In a one-to-one manner the operator A maps elements $f(t, a)$ of the space M into elements $F(x, \alpha)$ of the space N .

Let the following measurements of the function $F(x)$ be taken:

$$(y_1, x_1), \dots, (y_\ell, x_\ell). \quad (\text{A6.2})$$

The pair (x_i, y_i) denotes that the measured value of the function $F(x_i)$ at point x_i is y_i . We consider the following model of measuring:

1. The values of function $F(x)$ are measured with an additive error which does not depend on x :

$$y_i = F(x_i) + \xi_i, \\ E\xi = 0, \quad E\xi^2 = \sigma^2 < \infty.$$

2. The points x_i at which the measurements are taken are chosen randomly and independently according to some known nonvanishing probability measure $P(x)$. Below without loss in generality we consider the uniform distribution on $[a,b]$.

Given the operator A and the measurements (A6.2), it is required to estimate the function $f(t) = f(t, \alpha_0)$ in the set $f(t, \alpha)$. Here it is assumed that the problem of solving the operator equation (A6.1) may be ill-posed.

We call the problem of solving the operator equation (A6.1) on the basis of data obtained in accordance with the described model *the problem of estimating functions on the basis of indirect measurements*.

Note that the operator A maps a set of functions $f(t, \alpha), \alpha \in A$ into a set of functions

$$F(x, \alpha) = Af(t, \alpha).$$

Therefore any function $f(t, \alpha)$ has the image in the space \mathcal{N} . The solution $f(t, \alpha_0)$ of Eq. (A6.1) is the preimage in M of the regression $F(x, \alpha_0)$ in the space \mathcal{N} .

Since $f(t, \alpha_0) = f(t)$ belongs to the set $f(t, \alpha), \alpha \in A$, the preimage of the point that minimizes the functional

$$\begin{aligned} R(\alpha) &= \int (y - F(x, \alpha))^2 dP(x, y) = \int (y - Af(t, \alpha))^2 dP(x, y) \\ &= \sigma^2 + \int_a^b (Af(t, \alpha) - Af(t, \alpha_0))^2 dx \end{aligned} \quad (\text{A6.3})$$

is the solution of Eq. (A6.1).

However, it is impossible to obtain the exact minima (A6.3) using a finite number of measurements. One can only hope to obtain a function $F(x, \alpha^*)$ which is close (in the metric of space \mathcal{N}) to the regression, and then to choose as a solution of Eq. (A6.1) the preimage $f(t, \alpha^*)$ of this function in the space M .

Such an approach is not always successful: It is inconsistent if Eq. (A6.1) defines an ill-posed problem. In this case, widely different preimages in M may (though not necessarily) correspond to close images in \mathcal{N} .

In our case it implies that not all methods of risk minimization in the space of images may be utilized for solving the problem of estimating the results of indirect experiments, and there may exist a method for risk minimization which produces only those elements $F(x, \alpha^*)$ in the space \mathcal{N} whose preimages are close to the desired solution. These methods for risk minimization (if

they exist) should be utilized for solving ill-posed problems of estimating the results of indirect measurements.

The following text shows that under certain conditions the method of structural risk minimization may be utilized for solving ill-posed measurement problems. We shall prove that as the number of measurements increases, a sequence of solutions obtained by using the method of structural risk minimization converges to the desired function $f(t)$.

A6.2 THEOREMS ON ESTIMATING FUNCTIONS USING INDIRECT MEASUREMENTS

Consider a linear, completely continuous operator \mathbf{A} acting from the space L_2 into the space C , and let \mathbf{A}^* be the conjugate operator of \mathbf{A} . Then the operator $\mathbf{A}^*\mathbf{A}$ is also completely continuous. Let

$$\lambda_1^2 \geq \lambda_2^2 \geq \cdots \geq \lambda_m^2 \geq \cdots$$

be a complete system of its eigenvalues and let

$$\phi_1(t), \dots, \phi_m(t), \dots$$

be a complete orthonormal system of its eigenfunctions.

Consider also operator $\mathbf{A}\mathbf{A}^*$. It has the same set of eigenvalues, to which a complete orthonormal system of eigenfunctions

$$\psi_1(x), \dots, \psi_m(x), \dots$$

corresponds. Elements of ϕ_k and ψ_k satisfy the relations

$$\begin{aligned} A\phi_p(t) &= \lambda_p \psi_p(x), & p = 1, 2, \dots, \\ A^*\psi_p(x) &= \lambda_p \phi_p(t), & p = 1, 2, \dots \end{aligned} \tag{A6.4}$$

A solution of the operator equation (A6.1) can be expanded in a series in the system of functions $\phi_p(t)$, $p = 1, \dots$:

$$f(t, \alpha_0) = \sum_{p=1}^{\infty} \alpha_0^p \phi_p(t). \tag{A6.5}$$

We shall consider the function

$$f(t, \alpha_\ell) = \sum_{p=1}^{n(\ell)} \alpha_\ell^p \phi_p(t) \tag{A6.6}$$

to be an approximation to the solution (A6.5). Here $n(\ell)$ is the number of terms in the expansion (to be determined below) and $\alpha_\ell = (\alpha_\ell^1, \dots, \alpha_\ell^{n(\ell)})$ is the vector of parameters which yields the minimum for the functional:

$$R_{\text{emp}}(\alpha) = \frac{1}{\ell} \sum_{j=1}^{\ell} \left(y_j - \sum_{p=1}^{n(\ell)} \lambda_p \alpha^p \psi_p(x_j) \right)^2. \quad (\text{A6.7})$$

It turns out that under certain assumptions concerning the solution (A6.5) a function $n = n(\ell)$ exists such that as the sample size increases, the obtained approximations approach in probability the solution (A6.5) of the operator equation (A6.1).

The following three theorems are valid.

Theorem A6.1. *As ℓ increases, the sequence of approximations $f(t, \alpha_\ell)$ converges in probability to $f(t)$ in the metric L_2 , if the rule $n = n(\ell)$ for choosing the number of terms in expansion (A6.6) satisfies the conditions*

$$n(\ell) \xrightarrow[\ell \rightarrow \infty]{} \infty, \quad (\text{A6.8})$$

$$\frac{1}{\lambda_{n(\ell)}^2} \sqrt{\frac{n(\ell) \ln \ell}{\ell}} \xrightarrow[\ell \rightarrow \infty]{} 0. \quad (\text{A6.9})$$

To guarantee convergence in metric C, we make additional requirements.

Theorem A6.2. *Let the conjugate operator A^* be a bounded operator from space C into L_2 and the solution of operator equation (A6.1) be such that the condition*

$$\sup_t \left| \sum_{p=m}^{\infty} \alpha_0^p \phi_p(t) \right| = T(m), \quad T(m) \xrightarrow[m \rightarrow \infty]{} 0 \quad (\text{A6.10})$$

is fulfilled. Then the conditions (A6.8) and (A6.9) are sufficient to ensure convergence in probability of the functions $f(t, \alpha_\ell)$ to $f(t)$ in C metric.

Theorems A6.1 and A6.2 thus assert that if one approximates the solution of (A6.1) by means of an expansion in a finite number of eigenfunctions of the self-adjoint operator A^*A , then under appropriate choice of a number of terms in the expansion (satisfying the conditions (A6.8) and (A6.9)) the method of minimizing empirical risk (A6.7) ensures the convergence in probability of the obtained solutions to the desired one.

Now consider a structure in which element S_n is a set of functions that are expanded on the first n eigenfunctions. Let us denote by $|A|$ the norm of the

operator A acting from L_2 into C —that is, the smallest value $|A|$ for which the following inequality holds:

$$|F(x)|_C \leq |A| |f(t)|_{L_2},$$

where $|F(x)|_C$ is the norm of function $F(x)$ in the space \mathcal{N} and $|f(t)|_{L_2}$ is the norm of function $f(t)$ in the space \mathbf{M} . In the next theorem we show that if noise is such that

$$\frac{\sqrt{E\xi^{2p}}}{E\xi^2} = \tau < \infty, \quad p > 2, \quad (\text{A6.11})$$

then with probability $1 - \eta$ for sufficiently large n the inequality

$$R(\alpha_\ell^n) \leq \frac{R_{\text{emp}}(\alpha_\ell^n)}{\left(1 - 2\tau_n a(p) \sqrt{\frac{n \left(\ln \frac{2\ell}{n} + 1 \right) - \ln \eta}{\ell}}\right)_+} \quad (\text{A6.12})$$

holds, where

$$\tau_n = \frac{2|A|^2 \tau}{\lambda_n^2}.$$

The goal is to prove that by choosing the number n of an element S_n which contains the function that minimizes the right-hand side of inequality (A6.12) under constraint

$$n = n(\ell) \leq \frac{\ell}{\ln^2 \ell},$$

one satisfies conditions (A6.8) and (A6.9).

In other words, we will show that the standard procedure of the method of structural risk minimization leads to the construction of a sequence of functions that converges to the solution of the operator equation (A6.1).

Theorem A6.3. *Let a solution of the operator equation (A6.1) belong to the space L_2 and let the condition (A6.11) be satisfied. Then with probability $1 - \eta$ the bound (A6.12) is valid and the SRM principle based on this bound specifies a (random) sequence*

$$n = n(\ell)$$

such that for

$$n^*(\ell) = \min \left(n(\ell), \frac{\ell}{\ln^2 \ell} \right)$$

the following conditions are satisfied

$$\begin{aligned} n^*(\ell) &\xrightarrow[\ell \rightarrow \infty]{} \infty, \\ \frac{1}{\lambda_{n^*(\ell)}^2} \sqrt{\frac{n^*(\ell) \ln \ell}{\ell}} &\xrightarrow[\ell \rightarrow \infty]{} 0. \end{aligned} \quad (\text{A6.13})$$

Thus Theorems **A6.1** and **A6.2** point to a class of methods that ensures convergence of the sequence of obtained functions to the solution of the operator equation, while Theorem **A6.3** asserts that the SRM methods belongs to this class.

A6.3 PROOFS OF THE THEOREMS

A6.3.1 Proof of Theorem A6.1

The proof of this theorem is based on the technique that was developed for proving Theorem **A6.2**.

Let the conditions of Theorem **A6.1** be satisfied. Denote by

$$f(t, \alpha_\ell) = \sum_{k=1}^{n(\ell)} \alpha_\ell^k \phi_k(t)$$

the preimage of the function

$$F(x, \alpha_\ell) = \sum_{k=1}^{n(\ell)} \lambda_k \alpha_\ell^k \psi_p(x),$$

which minimizes the value of the empirical risk

$$R_{\text{emp}}(\alpha) = \frac{1}{\ell} \sum_{j=1}^{\ell} \left(y_j - \sum_{k=1}^{n(\ell)} \lambda_k \alpha^k \psi_k(x_j) \right)^2. \quad (\text{A6.14})$$

Our goal is to prove that $f(t, \alpha_\ell)$ converges in probability to

$$f(t) = \sum_{k=1}^{\infty} \alpha_0^k \phi_k(t)$$

in the metric L_2 , or equivalently that the sequence of random variables

$$v(\ell) = \int \left(\sum_{k=1}^{n(\ell)} \alpha_\ell^k \phi_k(t) - \sum_{k=1}^{\infty} \alpha_0^k \phi_k(t) \right)^2 dt \quad (\text{A6.15})$$

converges in probability to zero as ℓ increases.

Note that since the basis functions are orthonormal the following equality holds:

$$v(\ell) = \sum_{k=1}^{n(\ell)} \beta_k^2 + \sum_{k=n(\ell)+1}^{\infty} (\alpha_0^k)^2 = T_1(n(\ell)) + T_2(n(\ell)),$$

where $\beta_k = \alpha_\ell^k - \alpha_0^k$.

Since the solution belongs to L_2 , the sequence $T_2(n(\ell))$ tends to zero as $n(\ell)$ increases. Therefore to prove the theorem it is sufficient to show that

$$T_1(n(\ell)) \xrightarrow{\ell \rightarrow \infty} 0.$$

We bound the quantity

$$T_1(n(\ell)) = \sum_{k=1}^{n(\ell)} \beta_k^2. \quad (\text{A6.16})$$

Let $\alpha_\ell = (\alpha_\ell^1, \dots, \alpha_\ell^{n(\ell)})$ be a vector which minimizes the empirical risk (A6.14). We then rewrite (A6.14) for

$$\beta_\ell = (\beta_\ell^1, \dots, \beta_\ell^{n(\ell)}) = (\alpha_\ell^1 - \alpha_0^1, \dots, \alpha_\ell^{n(\ell)} - \alpha_0^{n(\ell)})$$

in the form

$$R_{\text{emp}}(\beta_\ell) = \frac{1}{\ell} \sum_{i=1}^{\ell} \bar{y}_i - 2 \sum_{k=1}^{n(\ell)} \lambda_k \beta_\ell^k G_k + \sum_{k,q=1}^{n(\ell)} \lambda_k \beta_\ell^k \lambda_q \alpha_\ell^q \sum_{j=1}^{\ell} \frac{\psi_k(x_j) \psi_q(x_j)}{\ell}, \quad (\text{A6.17})$$

where

$$\begin{aligned} G_k &= \frac{1}{\ell} \sum_{j=1}^{\ell} \bar{y}_j \psi_k(x_j), \\ \bar{y}_j &= \xi_j + \sum_{k=n+1}^{\infty} \lambda_k \alpha_0^k \psi_k(x_j). \end{aligned}$$

Denote by K the covariance matrix with elements K_{kq} given by

$$K_{kq} = \frac{1}{\ell} \sum_{j=1}^{\ell} \psi_k(x_j) \psi_q(x_j)$$

and by G the n -dimensional vector with coordinates $G = (G_1, \dots, G_n)^T$. Then the vector $y = (\beta_\ell^1 \lambda_1, \dots, \beta_\ell^n \lambda_n)^T$ which yields the minimum for (A6.17) is given by

$$\gamma = K^{-1} G.$$

Therefore the bound

$$|\gamma|^2 = |K^{-1}G|^2 \leq |K^{-1}|^2 |G|^2 \quad (\text{A6.18})$$

is valid.

On the other hand the inequality

$$|\gamma|^2 = \sum_{k=1}^{n(\ell)} (\beta_k^k \lambda_k)^2 > \lambda_{n(\ell)}^2 \sum_{k=1}^{n(\ell)} (\beta_k^k)^2 = \lambda_{n(\ell)}^2 T_1(n(\ell)) \quad (\text{A6.19})$$

holds true. From the inequalities (A6.18) and (A6.19) we obtain

$$T_1(n(\ell)) < \frac{1}{\lambda_{n(\ell)}^2} |K^{-1}|^2 |G|^2.$$

Thus to prove the theorem it is sufficient to bound from above the norm of the matrix K^{-1} and the norm of the vector G.

Note that the norm of K does not exceed μ_{\max} , the largest eigenvalue of the matrix; also note that the norm of the matrix K^{-1} is bounded as follows:

$$|K^{-1}| \leq \frac{1}{\mu_{\min}},$$

where μ_{\min} is the smallest eigenvalue of the matrix K.

Therefore we obtained the following bound:

$$T_1(n(\ell)) < \frac{|G|^2}{\lambda_{n(\ell)}^2 \mu_{\min}^2}. \quad (\text{A6.20})$$

To bound μ_{\min} from below we consider the positive definite quadratic form

$$F_n(x, \gamma) = \left(\sum_{k=1}^n \gamma_k \psi_k(x) \right)^2,$$

which we shall examine in the domain

$$\sum_{k=1}^n \gamma_k^2 \leq 1. \quad (\text{A6.21})$$

Since any completely continuous operator A acting from L_2 into C is bounded, the inequality

$$\sup_x \left| \sum_{k=1}^n \gamma_k \psi_k(x) \right| \leq |A| \left| \sum_{k=1}^n \frac{\gamma_k}{\lambda_k} \phi_k(t) \right|_{L_2} < |A| \sqrt{\sum_{k=1}^n \left(\frac{\gamma_k}{\lambda_k} \right)^2} < \frac{|A|}{\lambda_n} \sqrt{\sum_{k=1}^n \gamma_k^2}$$

holds, which implies that in the domain (A6.21) the inequality

$$\sup_x \left| \sum_{k=1}^n \gamma_k \psi_k(x) \right| < \frac{|A|}{\lambda_n}$$

is valid.

Now consider the expression

$$\frac{1}{\ell} \sum_{i=1}^{\ell} F_n(x_i, \gamma) = \frac{1}{\ell} \sum_{i=1}^{\ell} \left(\sum_{k=1}^n \gamma_k \psi_k(x_i) \right)^2.$$

Observe that

$$\begin{aligned} EF_n(x, \gamma) &= \sum_{k=1}^n \gamma_k^2, \\ \frac{1}{\ell} \sum_{i=1}^{\ell} F_n(x_i, \gamma) &= \sum_{kq=1}^n \gamma_k \gamma_q K_{kq}. \end{aligned} \tag{A6.22}$$

Using a rotation transformation, we arrive at a new, twice orthogonal system of functions $\psi^*(x)_1, \dots, \psi_n^*(x)$ such that

$$\begin{aligned} EF_n(x, \gamma^*) &= \sum_{k=1}^n (\gamma_k^*)^2, \\ \frac{1}{\ell} \sum_{i=1}^{\ell} F_n(x_i, \gamma^*) &= \sum_{k=1}^n \mu_k (\gamma_k^*)^2, \end{aligned} \tag{A6.23}$$

where μ_1, \dots, μ_n are eigenvalues of the matrix K .

To bound the eigenvalues we utilize Theorem 5.1 on the uniform convergence of the means to their mathematical expectations for a class of bounded functions. Since the functions $F(x, y^*)$ for $|\gamma^*| \leq 1$ are bounded by the quantity $|A|^2/\lambda_n^2$, the inequality

$$\begin{aligned} P \left\{ \sup_{\gamma^*} \left| EF_n(x, \gamma^*) - \frac{1}{\ell} \sum_{i=1}^{\ell} F_n(x_i, \gamma^*) \right| > \varepsilon \frac{|A|^2}{\lambda_n^2} \right\} \\ < 4 \exp \left\{ \left(\frac{n(\ln \frac{2\ell}{n} + 1)}{\ell} - \left(\varepsilon - \frac{1}{\ell} \right)^2 \right) \ell \right\} \end{aligned}$$

is valid. Taking (A6.23) into account, we obtain

$$\begin{aligned} P \left\{ \sup_{\gamma^*} \left| \sum_{k=1}^n (\gamma_k^*)^2 (1 - \mu_k) \right| > \varepsilon \frac{|A|^2}{\lambda_n^2} \right\} \\ < 4 \exp \left\{ \left(\frac{n(\ln \frac{2\ell}{n} + 1)}{\ell} - \left(\varepsilon - \frac{1}{\ell} \right)^2 \right) \ell \right\} \end{aligned} \quad (\text{A6.24})$$

We shall require that

$$P \left\{ \sup_{\gamma^*} \left| \sum_{k=1}^{n(\ell)} (\gamma^*)^2 (1 - \mu_k) \right| > \varepsilon^* \right\} \leq \frac{4}{\ln \ell}.$$

This is satisfied for

$$\varepsilon^* \leq \mathcal{E}(n, \ell) = \frac{|A|^2}{\lambda_n^2} \left(\sqrt{\frac{n \left(\frac{2\ell}{n} + 1 \right) + \ln \ln \ell}{\ell}} + \frac{1}{\ell} \right). \quad (\text{A6.25})$$

It follows from (A6.24) that with probability $1 - 4/\ln \ell$ all eigenvalues μ_1, \dots, μ_n are located in the interval

$$1 - \mathcal{E}(n, \ell) \leq \mu_k \leq 1 + \mathcal{E}(n, \ell). \quad (\text{A6.26})$$

This implies that with probability $1 - 4/\ln \ell$ the inequality

$$\min_k \mu_k > \max(1 - \mathcal{E}(n, \ell), 0) \quad (\text{A6.27})$$

is fulfilled. Substituting (A6.27) into (A6.20), we obtain that the inequality

$$T_1(n) < \frac{|G|^2}{\lambda_n^2 \left(1 - \frac{2|A|^2}{\lambda_n^2} \sqrt{\frac{n \left(\frac{2\ell}{n} + 1 \right) + \ln \ln \ell}{\ell}} \right)_+^2} \quad (\text{A6.28})$$

is valid with probability $1 - 4/\ln \ell$.

It remains to bound the quantity $|G|^2$:

$$|G|^2 = \sum_{k=1}^n G_k^2 = \sum_{k=1}^n \frac{1}{\ell^2} \left(\sum_{i=1}^{\ell} \bar{y}_i \psi_k(x_i) \right)^2.$$

For this purpose we compute the expectation

$$E|G|^2 = E \sum_{k=1}^n G_k^2 \leq \frac{\sigma^2 + T_2(0)}{\ell} n = C \frac{n}{\ell},$$

where C is a constant that does not depend on ℓ and n . To bound $|G|$ we utilize Chebyshev's inequality for the first moment of a positive random variable ξ

$$P\{\xi > \varepsilon\} < \frac{E\xi}{\varepsilon},$$

where $\varepsilon = (Cn \ln \ell)/\ell$. Since $E|G|^2 < Cn/\ell$, we obtain

$$P\left\{|G|^2 > \frac{Cn \ln \ell}{\ell}\right\} < \frac{1}{\ln \ell}.$$

Thus with probability $1 - 1/\ln \ell$,

$$|G|^2 \leq \frac{Cn \ln \ell}{\ell} \quad (\text{A6.29})$$

When we substitute (A6.29) into (A6.28), we obtain that for ℓ sufficiently large the inequality

$$T_1(n) < \frac{Cn \ln \ell}{\ell \lambda_n^2 \left(1 - \frac{|A|\tau}{\lambda_n^2} \sqrt{\frac{n \ln \ell}{\ell}}\right)_+^2}$$

is fulfilled with probability $1 - 5/\ln \ell$, where C is constant. Inequality (A6.29) implies that $T_1(n(\ell))$ tends to zero in probability as

$$\frac{1}{\lambda_{n(\ell)}^2} \sqrt{\frac{n(\ell) \ln \ell}{\ell}} \xrightarrow{\ell \rightarrow \infty} 0.$$

The theorem is proved.

A6.3.2 Proof of Theorem A6.2

Now let additionally the operator A^* be bounded from space L_2 into space C :

$$|f(t, \alpha)|_C \leq |A^*| |F(x, \alpha)|_{L_2}$$

and let the solution of the operator equation (A6.1) obey the additional restriction

$$\sup_t \left| \sum_{k=n+1}^{\infty} \alpha_0^k \phi_k(t) \right| \xrightarrow{\ell \rightarrow \infty} 0. \quad (\text{A6.30})$$

We show that in this case the conditions

$$n(\ell) \xrightarrow[\ell \rightarrow \infty]{} \infty, \quad (\text{A6.31})$$

$$\frac{1}{\lambda_{n(\ell)}^2} \sqrt{\frac{n(\ell) \ln \ell}{\ell}} \xrightarrow[\ell \rightarrow \infty]{} 0 \quad (\text{A6.32})$$

are sufficient so that the sequence of solutions $f(t, \alpha_\ell)$ can converge in probability to the solution of the operator equation **(A6.1)** in the metric C.

We use the notation

$$v(\ell) = \sup_t \left| \sum_{k=1}^{\infty} \alpha_0^k \phi_k(t) - \sum_{k=1}^{n(\ell)} \alpha_\ell^k \phi_k(t) \right|,$$

where $\alpha_\ell = (\alpha_\ell^1, \dots, \alpha_\ell^{n(\ell)})^T$ is the vector that yields the minimal value for **(A6.14)**. Our purpose is to prove that

$$v(\ell) \xrightarrow[\ell \rightarrow \infty]{P} 0.$$

Observe that

$$v(\ell) \leq \sup_t \left| \sum_{k=1}^{n(\ell)} \beta_k \phi_k(t) \right| + \sup_t \left| \sum_{k=n(\ell)+1}^{\infty} \alpha_0^k \phi_k(t) \right|, \quad (\text{A6.33})$$

where $\beta_k = \alpha_\ell^k - \alpha_0^k$.

In view of the condition **(A6.30)** of the theorem, the second summand in **(A6.33)** tends to zero with increasing ℓ . It is therefore sufficient to verify that

$$T_3(n(\ell)) = \sup_t \left| \sum_{k=1}^{n(\ell)} \beta_k \phi_k(t) \right| \xrightarrow[\ell \rightarrow \infty]{P} 0. \quad (\text{A6.34})$$

To prove this we shall use the bound

$$T_3^2(n(\ell)) < \frac{|A^*|^2}{\lambda_n^2} \sum_{k=1}^n \beta_k^2, \quad (\text{A6.35})$$

which is valid because the operator A^* is bounded.

In the course of the proof of Theorem **A6.1** it was shown that the bound

$$T_1(n(\ell)) = \sum_{p=1}^n \beta_p^2 < C \frac{n \ln \ell}{\ell \lambda_n^2 \left(1 - \frac{2|A|^2 \tau}{\lambda_n^2} \sqrt{\frac{n(\ln 2\ell/n + 1) + \ln \ln \ell}{\ell}} \right)_+^2}$$

holds with probability $1 - 4/\ln \ell$. Substituting this bound into (A6.35), we obtain that with probability $1 - 4/\ln \ell$ the inequality

$$T_3^2(\ell) < \frac{\frac{|A^*|^2 C}{\lambda_{n(\ell)}^4} \frac{n \ln \ell}{\ell}}{\left(1 - \frac{2|A|^2 \tau}{\lambda_n^2} \sqrt{\frac{n \ln \ell}{\ell}}\right)_+^2}$$

is satisfied. This bound implies that $T_3(\ell)$ approaches zero in probability provided that

$$\frac{1}{\lambda_{n(\ell)}^2} \sqrt{\frac{n(\ell) \ln \ell}{\ell}} \xrightarrow[\ell \rightarrow \infty]{} 0.$$

Theorem A6.2 is thus proved.

A6.3.3 Proof of Theorem A6.3

Let the number $n(\ell)$ of terms in the expansion of the solution of an operator equation satisfy the condition

$$n(\ell) \leq \frac{\ell}{\ln^2 \ell}$$

and be determined by the minimum value on the bounds of the risk (A6.12) over the elements of the structure. We show that if the solution of the operator equation (A6.1) satisfies

$$\left| \sum_{k=1}^{\infty} \alpha_0^k \phi_p(t) \right|_{L_2} = \sum_{k=1}^{\infty} (\alpha_0^k)^2 < \infty \quad (\text{A6.36})$$

and the errors of measurements are such that (A6.11) is valid, then the algorithm that chooses the element of the structure by minimizing (A6.12) satisfies the conditions

$$n(\ell) \xrightarrow[\ell \rightarrow \infty]{} \infty, \quad (\text{A6.37})$$

$$\frac{1}{\lambda_{n(\ell)}^2} \sqrt{\frac{n(\ell) \ln \ell}{\ell}} \xrightarrow[\ell \rightarrow \infty]{} 0. \quad (\text{A6.38})$$

To prove the theorem we need the following lemma.

Lemma. *Let the noise of the measurements be such that the inequality*

$$\frac{\sqrt[p]{E\xi^2 p}}{E\xi^2} < \tau, \quad p > 2 \quad (\text{A6.39})$$

holds true. Then for a set of functions satisfying the conditions

$$\left| \sum_{p=1}^{\infty} \alpha_0^p \phi_p(t) \right|_{L_2} = \sum_{k=1}^{\infty} (\alpha_0^k)^2 < \infty \quad (\text{A6.40})$$

and for a sufficiently large n the inequality

$$\tau_n = \sup_{\alpha} \frac{\sqrt[p]{E \left(y - A \left\{ \sum_{k=1}^{n(\ell)} \alpha^k \phi_k(t) \right\} \right)^{2p}}}{E \left(y - A \left\{ \sum_{k=1}^{n(\ell)} \alpha^k \phi_k(t) \right\} \right)^2} < \frac{2|A|^2 \tau}{\lambda_n^2}, \quad p > 2 \quad (\text{A6.41})$$

holds, where $|A|$ is the norm of the operator \mathbf{A} from L_2 to C .

Proof Recall that

$$y = F(x, \alpha_0) + \xi = \sum_{k=1}^{\infty} \lambda_k \alpha_0^k \psi_k(x) + \xi.$$

Therefore

$$\tau_n = \sup_{\alpha} \frac{\sqrt[p]{E \left(\xi + \delta(n, x) - \sum_{k=1}^{n(\ell)} \lambda_k \beta_k \psi_k(x) \right)^{2p}}}{E \left(\xi + \delta(n, x) - \sum_{k=1}^{n(\ell)} \lambda_k \beta_k \psi_k(x) \right)^2}, \quad (\text{A6.42})$$

where

$$\begin{aligned} \beta_p &= \alpha^p - \alpha_0^p, \\ \delta(n, x) &= \sum_{k=n(\ell)+1}^{\infty} \lambda_k \alpha_0^k \psi_k(x). \end{aligned}$$

We shall bound separately the denominator and numerator of the right-hand side of (A6.42):

$$E \left(\xi + \delta(n, x) - \sum_{k=1}^{n(\ell)} \lambda_k \beta_k \psi_k(x) \right)^2 = \sigma^2 + B^2 + \delta_n^2, \quad (\text{A6.43})$$

where

$$\begin{aligned} \sigma^2 &= E\xi^2 \\ B^2 &= \sum_{k=1}^{n(\ell)} \lambda_k^2 \beta_k^2, \\ \delta_n^2 &= \sum_{k=n(\ell)+1}^{\infty} \lambda_k^2 (\alpha_0^k)^2 \xrightarrow[\ell \rightarrow \infty]{} 0. \end{aligned}$$

To bound the numerator we use the Minkowski inequality

$$\sqrt[p]{E|A+B|^p} \leq \sqrt[p]{E|A|^p} + \sqrt[p]{E|B|^p}.$$

We obtain

$$\begin{aligned} & \sqrt[p]{E \left((\xi + \delta(n, \ell)) - \sum_{k=1}^{n(\ell)} \lambda_k \beta_k \psi_k(x) \right)^{2p}} \\ & \leq \left(\sqrt[2p]{E\xi^{2p}} + \sqrt[2p]{E\delta^{2p}(n, \ell)} + \sqrt[2p]{E \left(\sum_{k=1}^{n(\ell)} \lambda_k \beta_k \psi_k(x) \right)^{2p}} \right)^2 \\ & \leq 3 \left(\sqrt[p]{E\xi^{2p}} + \sqrt[p]{E\delta^{2p}(n, \ell)} + \sqrt[p]{E \left(\sum_{k=1}^{n(\ell)} \lambda_k \beta_k \psi_k(x) \right)^{2p}} \right). \quad (\text{A6.44}) \end{aligned}$$

Since the desired function belongs to L_2 and an operator A acting from L_2 into C is bounded, we have

$$\begin{aligned} \sup_x |\delta(n, x)| & \leq |A| \sqrt{\sum_{k=n(\ell)+1}^{\infty} (\alpha_0^k)^2} \xrightarrow{\ell \rightarrow \infty} 0, \\ \sup_x \left| \sum_{k=1}^{n(\ell)} \lambda_k \beta_k \psi_k(x) \right| & \leq |A| \frac{B}{\lambda_n}. \end{aligned} \quad (\text{A6.45})$$

Substituting **(A6.45)** into **(A6.44)** we obtain for sufficiently large n the following bound for the numerator:

$$\sqrt[p]{E \left(y - \sum_{k=1}^{n(\ell)} \lambda_k \beta_k \psi_k(x) \right)^{2p}} \leq 3 \left(\sqrt[p]{E\xi^{2p}} + |A|^2 \frac{B^2}{\lambda_n^2} \right).$$

Substituting in **(A6.42)** bounds for the numerator and the denominator and taking into account that $\tau \geq 1$ and for large n we have $|A|/\lambda_n > 1$, one obtains

$$\frac{\sqrt[p]{E \left(y - \sum_{k=1}^{n(\ell)} \alpha^k \psi_k(t) \right)^{2p}}}{E \left(y - \sum_{k=1}^{n(\ell)} \alpha^k \psi_k(t) \right)^2} \leq 3 \frac{\sqrt[p]{E\xi^{2p}} + \frac{|A|^2 B^2}{\lambda_n^2}}{E\xi^2 + B^2} \leq 3 \frac{|A|^2 \tau}{\lambda_n^2}.$$

The lemma has been proved.

Based on the result of this lemma one can assert that for the function $F(x, \alpha_\ell^n)$ which minimizes the empirical risk (A6.7) in the set

$$F(x, \alpha) = \sum_{k=1}^n \lambda_k \alpha^k \psi_k(x)$$

with probability $1 - \eta$ for sufficiently large ℓ the inequality

$$R(\alpha_\ell^n) < \frac{\frac{1}{\ell} \sum_{i=1}^{\ell} \left(y_i - \sum_{k=1}^n \lambda_k \alpha_\ell^k \psi_k(x_i) \right)^2}{\left(1 - 4 \frac{|A|^2 \tau}{\lambda_n^2} a(p) \sqrt{\frac{n \left(\ln \frac{2\ell}{n} + 1 \right)}{\ell}} - \ln \eta / 4 \right)_+} \quad (\text{A6.46})$$

holds.

Now for a given set of ℓ observations we choose $n = n(\ell)$ that defines the function minimizing the right-hand side of this bound. We prove that the chosen $n = n(\ell)$ is such that expression (A6.37) and (A6.38) of convergence in probability are valid.

First we verify that (A6.37) is valid. Assume the contrary. Let $\alpha_0^n \neq 0$, $m < n$, but at the same time let the inequality

$$\begin{aligned} & \frac{\frac{1}{\ell} \sum_{i=1}^{\ell} \left(y_i - \sum_{k=1}^m \lambda_k \alpha_\ell^k \psi_k(x_i) \right)^2}{\left(1 - \frac{4|A|^2 \tau}{\lambda_m^2} a(p) \sqrt{\frac{m \left(\ln \frac{2\ell}{m} + 1 \right)}{\ell}} - \ln \eta \right)_+} \\ & < \frac{\frac{1}{\ell} \sum_{i=1}^{\ell} \left(y_i - \sum_{k=1}^n \lambda_k \alpha_\ell^k \psi_k(x_i) \right)^2}{\left(1 - \frac{4|A|^2 \tau}{\lambda_n^2} a(p) \sqrt{\frac{n \left(\ln \frac{2\ell}{n} + 1 \right)}{\ell}} - \ln \eta \right)_+} \end{aligned}$$

be fulfilled for any $\ell > \ell_0$. Represent the quantity $R(\alpha_\ell^m)$ in the form

$$\begin{aligned} R(\alpha_\ell^m) &= E \left(y - \sum_{k=1}^m \lambda_k \alpha_\ell^k \psi_k(x) \right)^2 \\ &= E \left(\xi + \delta(x, m) - \sum_{p=1}^m \lambda_p \beta_p \psi_p(x) \right)^2, \end{aligned}$$

where

$$\delta(x, m) = \sum_{k=m+1}^{\infty} \lambda_k \alpha_0^k \psi_k(x),$$

and bound this quantity from below:

$$R(\alpha_\ell^m) > R(\alpha_0) = \sigma^2 + \sum_{k=m+1}^{\infty} (\lambda_k \alpha_0^k)^2$$

Thus, the bound

$$\sigma^2 + \sum_{k=m+1}^{\infty} (\lambda_k \alpha_0^k)^2 < \frac{\frac{1}{\ell} \sum_{i=1}^{\ell} \left(y_i - \sum_{k=1}^n \lambda_k \alpha_\ell^k \psi_k(x_i) \right)^2}{\left(1 - \frac{4|A|^2 \tau}{\lambda_n^2} a(p) \sqrt{\frac{n \left(\ln \frac{2\ell}{n} + 1 \right)}{\ell}} - \ln \eta \right)_+} \quad (\text{A6.47})$$

is valid with probability at least $1 - \eta$. Now we transform and bound the expression appearing in the numerator on the right-hand side of (A6.47):

$$\begin{aligned} R_{\text{emp}}(\alpha_\ell) &= \frac{1}{\ell} \sum_{i=1}^{\ell} \left(y_i - \sum_{k=1}^n \lambda_k \alpha_\ell^k \psi_k(x_i) \right)^2 \\ &= \frac{1}{\ell} \sum_{i=1}^{\ell} \left(\xi_i + \delta(x_i, n) - \sum_{k=1}^n \lambda_k \beta_k \psi_k(x_i) \right)^2 \\ &< \frac{1}{\ell} \sum_{i=1}^{\ell} (\xi_i + \delta(x_i, n))^2 \\ &= \frac{1}{\ell} \sum_{i=1}^{\ell} \xi_i^2 + \frac{1}{\ell} \sum_{i=1}^{\ell} \delta^2(x_i, n) + \frac{2}{\ell} \sum_{i=1}^{\ell} \xi_i \delta(x_i, n). \end{aligned}$$

We obtained this inequality because minimum of empirical risk is achieved when $\beta_k = \alpha_\ell^k - \alpha_0^k$. Note that in view of the law of large numbers we obtain

$$\frac{1}{\ell} \sum_{i=1}^{\ell} \xi_i^2 \xrightarrow[\ell \rightarrow \infty]{} \sigma^2,$$

$$\frac{1}{\ell} \sum_{i=1}^{\ell} \xi_i \delta(x_i, n) \xrightarrow[\ell \rightarrow \infty]{} 0,$$

$$\frac{1}{\ell} \sum_{i=1}^{\ell} \delta^2(x_i, n) \xrightarrow[\ell \rightarrow \infty]{} \sum_{k=n+1}^{\infty} (\lambda_k \alpha_0^k)^2$$

Therefore the inequality

$$\sigma^2 + \sum_{k=m+1}^{\infty} (\lambda_k \alpha_0^k)^2 < \sigma^2 + \sum_{p=n+1}^{\infty} (\lambda_p \alpha_0^p)^2 \quad (\text{A6.48})$$

is satisfied with probability $1 - \eta$ for all sufficiently large ℓ .

However, for $m < n$ the inequality (A6.48) is obviously invalid with probability 1. The construction proves the validity of (A6.37).

Now we show that (A6.38) is also valid. For this purpose note that the inequalities

$$\begin{aligned} \frac{1}{\ell} \sum_{i=1}^{\ell} (\xi_i + \delta(x_i, n))^2 &> \min_{\alpha} R_{\text{emp}}(\alpha, n) \\ &> \min_{\alpha, \gamma} \frac{1}{\ell} \sum_{i=1}^{\ell} \left(\xi_i - \gamma \delta(x_i, n) - \sum_{p=1}^n \lambda_p \beta_p \psi_p(x_i) \right)^2 \end{aligned} \quad (\text{A6.49})$$

always hold. Compute the expectation of the left-hand side of inequality (A6.49):

$$E \left\{ \frac{1}{\ell} \sum_{i=1}^{\ell} (\xi_i + \delta(x_i, n))^2 \right\} = \sigma^2 + \sum_{k=n+1}^{\infty} (\lambda_k \alpha_0^k)^2 = \sigma^2 + T_2(n).$$

Observe that for a fixed n the relation

$$\frac{1}{\lambda_n^2} \sqrt{\frac{n \ln \ell}{\ell}} \xrightarrow[\ell \rightarrow \infty]{} 0$$

is valid. Therefore for fixed n the inequality

$$\lim_{\ell \rightarrow \infty} \frac{R_{\text{emp}}(\alpha_\ell^n)}{\left(1 - \frac{4|A|^2\tau}{\lambda_n^2}a(p)\sqrt{\frac{n\left(\ln \frac{2\ell}{n} + 1\right) - \ln \eta/4}{\ell}}\right)_+} < \sigma^2 + T_2(n) \quad (\text{A6.50})$$

is fulfilled.

Since the inequality (A6.50) is valid for any fixed n and the conditions

$$T_2(n(\ell)) \xrightarrow[\ell \rightarrow \infty]{} 0,$$

$$n(\ell) \xrightarrow[\ell \rightarrow \infty]{} \infty$$

are fulfilled, it follows that the inequality

$$\lim_{\ell \rightarrow \infty} \min_{n(\ell) < \ell / \ln^2 \ell} \left\{ \frac{R_{\text{emp}}(\alpha_\ell^{n(\ell)})}{\left(1 - \frac{4||A||^2\tau}{\lambda_{n(\ell)}^2}\sqrt{\frac{n(\ell)\left(\ln \frac{2\ell}{n} + 1\right) - \ln \eta/4}{\ell}}\right)_+} \right\} \leq \sigma^2 \quad (\text{A6.51})$$

holds true.

On the other hand we utilize the following bounds which will be derived as follows:

$$\begin{aligned} ER_{\text{emp}}(\alpha_\ell, \gamma_\ell, n(\ell)) &= E \min_{\alpha, \gamma} \frac{1}{\ell} \sum_{i=1}^{\ell} \left(\xi_i - \gamma \delta(x_i, n(\ell)) - \sum_{k=1}^{n(\ell)} \lambda_k \beta_k \psi_k(x_i) \right)^2 \\ &= \sigma^2 \left(1 - \frac{n(\ell) + 1}{\ell}\right), \end{aligned} \quad (\text{A6.52})$$

$$\begin{aligned} \text{Var}R_{\text{emp}}(\alpha_\ell, \gamma_\ell, n(\ell)) &= ER_{\text{emp}}^2(\alpha_\ell, \gamma_\ell, n(\ell)) - (ER_{\text{emp}}(\alpha_\ell, \gamma_\ell, n(\ell)))^2 \\ &\leq \frac{\tau^2 \sigma^4 + \sigma^4}{\ell^2} (n(\ell) + 1) = \Delta \frac{n(\ell)}{\ell^2}. \end{aligned} \quad (\text{A6.53})$$

Here α_ℓ and γ_ℓ are the values of the parameters which minimize $R_{\text{emp}}(\alpha, \gamma_\ell, n(\ell))$, and A is a constant that does not depend on ℓ . (We shall verify these bounds below.)

Now using Chebyshev inequality we obtain

$$P \left\{ \left| R_{\text{emp}}(\alpha_\ell, \gamma_\ell, n(\ell)) - \sigma^2 \left(1 - \frac{n(\ell) + 1}{\ell} \right) \right| > \varepsilon \right\} < \frac{\Delta n(\ell)}{\varepsilon^2 \ell^2}$$

Recall that we chose such $n(\ell)$ that $n(\ell) \leq \ell / \ln^2 \ell$. Therefore

$$\begin{aligned} \sum_{i=1}^{\infty} P \left\{ \left| R_{\text{emp}}(\alpha_\ell, \gamma_\ell, n(\ell)) - \sigma^2 \left(1 - \frac{n(\ell) + 1}{\ell^2} \right) \right| > \varepsilon \right\} \\ < \Delta \sum_{\ell=1}^{\infty} \frac{1}{\varepsilon^2 \ell \ln^2 \ell} < \infty \end{aligned}$$

and consequently the convergence

$$\lim_{\ell \rightarrow \infty} R_{\text{emp}}(\alpha_\ell, \gamma_\ell, n(\ell)) = \sigma^2 \quad (\text{A6.54})$$

is valid with probability one according to Borel–Cantelli lemma.

Therefore with probability one, the inequalities (A6.51) and (A6.54) hold. This implies that with probability one we obtain

$$\frac{1}{\lambda_{n(\ell)}^2} \sqrt{\frac{n(\ell) \ln \ell}{\ell}} \xrightarrow{\ell \rightarrow \infty} 0.$$

This expression constitutes the statement of Theorem A6.3.

In the proof of Theorem A6.3 we have used the equality (A6.52) and the inequality (A6.53). We shall now derive them. Consider the expression

$$ER_{\text{emp}}(\alpha_\ell, \gamma_\ell, n(\ell)) = E \min_{\alpha, \gamma} \frac{1}{\ell} \sum_{i=1}^{\ell} \left(\xi_i - \gamma \delta(x_i, n(\ell)) - \sum_{k=1}^{n(\ell)} \lambda_k \beta_k \psi_k(x_i) \right)^2$$

Using a rotation transformation we arrive at a coordinate system $\psi_1^*(x), \dots, \psi_{n+1}^*$ such that

$$\psi_k^*(x_i) \psi_q^*(x_i) = \begin{cases} \mu_p & \text{for } k = q, \\ 0 & \text{for } k \neq q. \end{cases}$$

In this coordinate system we obtain

$$R_{\text{emp}}(\alpha_\ell, \gamma_\ell, n) = \sum_{i=1}^{\ell} \xi_i^2 - \sum_{k=1}^{n+1} \frac{G_k^2}{\mu_k},$$

where

$$G_k = \sum_{i=1}^{\ell} \xi_i \psi_k^*(x_i)$$

We have thus obtained

$$ER_{\text{emp}}(\alpha_\ell, \gamma_\ell, n) = \sigma^2 - E \sum_{p=1}^{n+1} \sum_{i,j=1}^{\ell} \frac{\xi_i \xi_j \psi_p^*(x_i) \psi_p^*(x_j)}{\ell^2 \mu_p} = \sigma^2 \left(1 - \frac{n+1}{\ell}\right),$$

$$\begin{aligned} \text{Var}R_{\text{emp}}(\alpha_\ell, \gamma_\ell, n) &= ER_{\text{emp}}^2(\alpha_\ell, \gamma_\ell, n(\ell)) - (ER_{\text{emp}}(\alpha_\ell, \gamma, n(\ell)))^2 \\ &= \sum_{p=1}^{n+1} E \left(\frac{G_p^2}{\mu_p} \right)^2 - \left(E \frac{G_p^2}{\mu_p} \right)^2 \\ &\leq A \frac{n c l}{\ell^2}. \end{aligned}$$

The theorem is proved.

STOCHASTIC ILL-POSED PROBLEMS

In the Appendix to Chapter 6 we showed that the SRM principle can be used for solving special stochastic ill-posed problems, namely, for problems of interpreting the indirect measurements. In this chapter we consider stochastic ill-posed problems as a generalization of the classical ill-posed problems presented in Chapter 1 and in the Appendix to Chapter 1. To solve stochastic ill-posed problems, we utilize the same regularization method that we used in the Appendix to Chapter 1 for solving classical (deterministic) ill-posed problems.

Using the theory of stochastic ill-posed problems, we then consider the problem of density (conditional density) estimation. We construct different estimators of the density (conditional density) which include both classical nonparametric estimators and new ones.

7.1 STOCHASTIC ILL-POSED PROBLEMS

Consider the operator equation

$$Af = F \quad (7.1)$$

defined by the continuous operator A which maps in a one-to-one manner the elements f of a metric space E_1 into the elements F of a metric space E_2 .

Suppose that a solution $f \in \mathcal{B} \subset E_1$ of Eq. (7.1) exists and is unique, but unstable; that is, the problem of solving (7.1) is ill-posed (see Chapter 1 for the definition of ill-posed problems).

We would like to find the solution to this equation when instead of Eq. (7.1) we are given its approximations.

We will distinguish between two cases:

Case 1. The Equation with Approximately Defined Right-Hand Side. We consider the situation where instead of the right-hand side F of Eq. (7.1) we are given a sequence of random (determined by probability spaces $(\Omega_\ell, \mathcal{F}_\ell, P_\ell)$, $\ell = 1, 2, \dots$) functions

$$F_1, \dots, F_\ell, \dots \quad (7.2)$$

which converge in probability to the unknown function F .

Case 2. The Equation with Both the Operator and the Right-Hand Side Approximately Defined. We consider the situation where instead of Eq. (7.1) we are given both a random sequence of approximations (7.2) to the function F on the right-hand side of equality (7.1) and a random sequence of operators in (7.1)

$$A_1, \dots, A_\ell, \dots \quad (7.3)$$

(determined by probability spaces $(\tilde{\Omega}_\ell, \tilde{\mathcal{F}}_\ell, \tilde{P}_\ell)$, $\ell = 1, 2, \dots$), which converges in probability to the unknown operator A (the distance between two operators will be defined later).

In other words, the following schemes are considered: For any given ℓ there exists a set Ω_ℓ of random events $\omega \in \Omega_\ell$ such that any $\omega^* \in \Omega_\ell$ specifies a (random) function $F_\ell = F(x, \omega^*)$ (in the space E_2) and there exists a set $\tilde{\Omega}_\ell$ of random events such that any $\bar{\omega}^* \in \tilde{\Omega}_\ell$ specifies an operator $A_\ell = A(\bar{\omega}^*)$.

In the first case on the basis of the sequence of (random) functions (7.2) converging in probability (in metric $\rho_{E_2}(F, F_\ell)$ of the space E_2) to the unknown function F

$$\lim_{\ell \rightarrow \infty} P\{\rho(F, F_\ell) > \varepsilon\} = 0, \quad \forall \varepsilon > 0, \quad (7.4)$$

one has to find a sequence of functions

$$f_1, \dots, f_\ell, \dots$$

converging in probability (in the metric $\rho_{E_1}(f, f_\ell)$ of the space E_1) to the solution of Eq. (7.1).

In the second case, on the basis of the both sequence of functions (7.2) converging in probability to F and the sequence of operators (7.3) converging in probability (for a given distance $\|A_\ell - A\|$) to operator A of Eq. (7.1)

$$\lim_{\ell \rightarrow \infty} P\{\|A_\ell - A\| > \delta\} = 0, \quad \forall \delta > 0, \quad (7.5)$$

one has to find a sequence of functions

$$f_1, \dots, f_\ell, \dots$$

converging in probability (in the metric $\rho_{E_1}(f, f_\ell)$ of the space E_1) to the solution of Eq. (7.1).

We call the problem of solving Eq. (7.1) on the basis of random sequence (7.2) a stochastic ill-posed problem with approximations on the right-hand side.

We call the problem of solving Eq. (7.1) on the basis of both random sequence (7.2) and random sequence (7.3) a stochastic ill-posed problem with approximations on the right-hand side and approximations on the operator.

Example 1: Problem of Density Estimation. In Chapter 1 we considered the problem of density estimation as a problem of solving the equation

$$\int_{-\infty}^x p(t) dt = F(x), \quad (7.6)$$

where the distribution function $F(x)$ is unknown, but we are given the data

$$x_1, \dots, X_f, \dots$$

In this example the space Ω_ℓ of random events is determined by the space of samples $w = x_1, \dots, x_\ell$ of size ℓ . The random functions can be determined by w as follows:

$$F_\ell(x, \omega) = \frac{1}{\ell} \sum_{i=1}^{\ell} \theta(x - x_i). \quad (7.7)$$

It is known (Glivenko–Cantelli theorem) that for these functions convergence (7.4) takes place, where

$$\rho(F(x), F_\ell(x)) = \sup_x |F(x) - F_\ell(x)|.$$

Therefore the problem of density estimation is the following: On the basis of the (random) approximations F_ℓ , $\ell = 1, 2, \dots$ (on the basis of observations $x_1, \dots, x_\ell, \dots$), find a sequence of functions f_ℓ , $\ell = 1, 2, \dots$, that converges in probability to the solution of Eq. (7.1).

Example 2: Problem of Ratio of Densities Estimation. The optimal (Bayesian) decision rule in the pattern recognition problem has the form

$$r(x) = \theta \left\{ \frac{p_1(x)}{p_2(x)} - \frac{p_1}{1-p_1} \right\},$$

where $p_1(x)$ and $p_2(x)$ are the probability densities of vectors of the two classes in the X space, and p_1 is probability of occurrence of the vectors of the first class. Therefore to estimate from the training data the optimal decision rule, one has to estimate from the data the value p_1 (which is easy) and the function

$$T(x) = \frac{p_1(x)}{p_2(x)}$$

determining the ratio of the densities.

From the formal point of view to estimate the function $T(x)$ from the training set means to solve the integral equation

$$\int_{-\infty}^x T(u) dF^{(2)}(u) = F^{(1)}(x) \quad (7.8)$$

in the situation where the distribution functions $F^{(2)}(x)$ and $F^{(1)}(x)$ of the vectors belonging to the different classes are unknown, but we are given examples

$$x_1, \dots, x_\ell$$

containing a examples

$$x_1, \dots, x_a$$

belonging to the first class and containing b examples

$$\bar{x}_1, \dots, \bar{x}_b$$

belonging to the second class.

In this example the space Ω_ℓ of random events is determined by the space of samples $\omega = x_1, \dots, x_a, \bar{x}_1, \dots, \bar{x}_b$ of size $\ell = a + b$.

The random functions $F_\ell^{(1)}(x)$ and $F_\ell^{(2)}(x)$ are determined by the event ω (sample x_1, \dots, x_ℓ) as follows:

$$\begin{aligned} F_\ell^{(1)}(x, \omega) &= \frac{1}{a} \sum_{i=1}^a \theta(x - x_i), \\ F_\ell^{(2)}(x, \omega) &= \frac{1}{b} \sum_{i=1}^b \theta(x - \bar{x}_i). \end{aligned} \quad (7.9)$$

Thus the problem of estimating the ratio of densities is the problem of solving the integral equation (7.8) where instead of the right-hand side $F^{(1)}(x)$ of the approximation (7.9) is given and instead of exact operator which is determined by the function $F^{(2)}(x)$ the approximation to operator which is determined by the function (7.9) is given.

7.2 REGULARIZATION METHOD FOR SOLVING STOCHASTIC ILL-POSED PROBLEMS

For solving stochastic ill-posed problems we use the same regularization method considered in the Appendix to Chapter 1 for solving deterministic ill-posed problems.

Below we, once more, describe this method. Consider a lower semi-continuous functional $W(f)$ satisfying the following properties:

1. The solution of Eq. (7.1) belongs to D , the domain of definition of the functional $W(f)$.
2. The functional $W(f)$ takes on real nonnegative values in D .
3. The sets $\mathcal{M}_c = \{f : W(f) \leq c\}$ are compact in E_1 .

Let us construct the functional

$$R_{\gamma_\ell}(f, F_\ell) = \rho_{E_2}^2(Af, F_\ell) + \gamma_\ell W(f),$$

where $\rho_{E_2}(\cdot, \cdot)$ is a metric in the space E_2 , $F_\ell = F_\ell(x)$ is a random function, and $\gamma_\ell > 0$ is a parameter of regularization. Let the function f_ℓ provide the minimum to this functional.

Below we consider the case where

$$\gamma_\ell \rightarrow 0, \quad \text{as } \ell \rightarrow \infty.$$

Under these conditions the following theorems describing the relationship between the distributions of two random variables—namely, random variable $\rho_{E_2}(F, F_\ell)$ and random variable $\rho_{E_1}(f, f_\ell)$ —hold true.

Theorem 7.1. *For any positive ε and μ there exists a positive number $n(\varepsilon, \mu)$ such that for all $\ell > n(\varepsilon, \mu)$ the inequality*

$$P\{\rho_{E_1}(f_\ell, f) > \varepsilon\} \leq P\{\rho_{E_2}(F_\ell, F) > \sqrt{\gamma_\ell \mu}\} \quad (7.10)$$

is fulfilled.

For the case where E_1 is a Hilbert space the following theorem holds true.

Theorem 7.2. *Let E_1 be a Hilbert space, A in (7.1) be a linear operator and*

$$W(f) = \||f||^2 = \int f^2(t) dt.$$

Then for any positive ε there exists a number $n(\varepsilon)$ such that for all $\ell > n(\varepsilon)$ the inequality

$$P\{||f_\ell - f||^2 > \varepsilon\} < 2P\{\rho_{E_2}^2(F_\ell, F) > \frac{\varepsilon}{2}\gamma_\ell\}$$

is fulfilled.

These theorems are a generalization of Theorems A1.1 and A1.2 for the stochastic case (see Appendix to Chapter 1).

Corollary. *From Theorems 7.1 and 7.2 it follows that if approximations F_ℓ on the right-hand side of the operator equation (7.1) converge to the true function $F(x)$ in the metric of space E_2 with the rate $r(\ell)$, then the sequence of the solutions to the equation converges in probability to the desired one if*

$$\frac{r(\ell)}{\sqrt{\gamma_\ell}} \xrightarrow[\ell \rightarrow \infty]{} 0$$

and γ_ℓ converges to zero with $\ell \rightarrow \infty$.

This chapter also considers a problem of solving the operator equation

$$Af = F$$

under the condition where (random) approximations are given not only for the function on the right-hand side of the equation, but for the operator as well. We will assume that instead of the exact operator A we are given a sequence of approximations A_ℓ , $\ell = 1, 2, \dots$, defined by a sequence of random continuous operators which converge in probability (below we will specify the definition of closeness of two operators) to operator A .

As before, we consider the problem of solving the operator equation by a regularization method—that is, by minimizing the functional

$$R_{\gamma_\ell}^*(f, F_\ell, A_\ell) = \rho_{E_2}^2(A_\ell f, F_\ell) + \gamma_\ell W(f). \quad (7.11)$$

(Under the considered requirements to operator A_ℓ and function $F_\ell \in E_2$ there exists the minimum of this functional. The uniqueness of minima is not necessary.)

We will measure the closeness of operator A_ℓ and operator A , by the distance

$$\|A_\ell - A\| = \sup_{f \in D} \frac{\|A_\ell f - Af\|_{E_2}}{W^{1/2}(f)} \quad (7.12)$$

Theorem 7.3. *For any $\varepsilon > 0$ and any constant $C_1, C_2 > 0$ there exists a value $\gamma_0 > 0$ such that for any $\gamma_\ell \leq \gamma_0$ the inequality*

$$P\{\rho_{E_1}(f_\ell, f) > \varepsilon\} \leq P\{\rho_{E_2}(F_\ell, F) > C_1\sqrt{\gamma_\ell}\} + P\{\|A_\ell - A\| > C_2\sqrt{\gamma_\ell}\} \quad (7.13)$$

holds true.

Corollary. *From this theorem it follows that if the approximations F_ℓ on the right-hand side of the operator equation converge to the true function $F(x)$ in the metric of the space E_2 with the rate of convergence $r(\ell)$, and the approximations A_ℓ converge to the true operator A in the metric defined in Eq. (7.12) with the rate of convergence $r_A(\ell)$, then there exists a function*

$$r_0(\ell) = \max\{r(\ell), r_A(\ell)\} \xrightarrow[\ell \rightarrow \infty]{} 0$$

such that the sequence of solutions to the equation converges in probability to the desired one if

$$\frac{r_0(\ell)}{\sqrt{\gamma_\ell}} \xrightarrow[\ell \rightarrow \infty]{} 0$$

and γ_ℓ converges to zero with $\ell \rightarrow \infty$.

7.3 PROOFS OF THE THEOREMS

7.3.1 Proof of Theorem 7.1

By definition, for any ℓ the chain of inequalities[†]

$$\begin{aligned} \gamma_\ell W(f_\ell) &\leq R(f_\ell, F_\ell) \leq R(f, F_\ell) \\ &= \rho_2^2(Af, F_\ell) + \gamma_\ell W(f) = \rho_2^2(F, F_\ell) + \gamma_\ell W(f) \end{aligned} \quad (7.14)$$

is valid, where f_ℓ is the function that minimizes $R(f, F_\ell)$. Therefore the inequality

$$W(f_\ell) \leq W(f) + \frac{\rho_2^2(F_\ell, F)}{\gamma_\ell} \quad (7.15)$$

is valid. Moreover, clearly

$$\rho_2^2(Af_\ell, F_\ell) \leq R(f_\ell, F_\ell). \quad (7.16)$$

[†]Here and below we set $\rho_{E_i} = \rho_i$ for notational simplicity.

Utilizing the triangle inequality and the bounds (7.14) and (7.16), we obtain the inequalities

$$\begin{aligned}\rho_2(Af_\ell, F) &\leq \rho_2(Af_\ell, F_\ell) + \rho_2(F_\ell, F) \\ &\leq \rho_2(F_\ell, F) + \sqrt{\rho_2^2(F_\ell, F) + \gamma_\ell W(f)}.\end{aligned}\quad (7.17)$$

Furthermore, for any $\varepsilon > 0$ and $C > W(f)$ the equality

$$\begin{aligned}P\{\rho_1(f_\ell, f) \leq \varepsilon\} &= P\left\{\rho_1(f_\ell, f) \leq \varepsilon | W(f) + \frac{\rho_2^2(F_\ell, F)}{\gamma_\ell} \leq C\right\} P\left\{W(f) + \frac{\rho_2^2(F_\ell, F)}{\gamma_\ell} \leq C\right\} \\ &+ P\left\{\rho_1(f_\ell, f) \leq \varepsilon | W(f) + \frac{\rho_2^2(F_\ell, F)}{\gamma_\ell} > C\right\} P\left\{W(f) + \frac{\rho_2^2(F_\ell, F)}{\gamma_\ell} > C\right\}\end{aligned}\quad (7.18)$$

is valid. Now let the condition

$$W(f) + \frac{\rho_2^2(F_\ell, F)}{\gamma_\ell} \leq C$$

be fulfilled. Then it follows from (7.15) that the inequality

$$W(f_\ell) \leq C$$

is valid; that is, f_ℓ belongs to a compactum. In view of the lemma on the continuity of the inverse operator A^{-1} to the operator \mathbf{A} on a compactum (see Appendix to Chapter 1), we obtain that there exists a 6 such that inequality

$$\rho_1(f_\ell, f) \leq \varepsilon$$

is fulfilled as long as inequality

$$\rho_2(Af_\ell, F) \leq \delta$$

is fulfilled. Hence we have for sufficiently large ℓ that

$$\begin{aligned}P\left\{\rho_1(f_\ell, f) \leq \varepsilon | W(f) + \frac{\rho_2^2(F_\ell, F)}{\gamma_\ell} \leq C\right\} \\ \geq P\left\{\rho_2(Af_\ell, F) \leq \delta | W(f) + \frac{\rho_2^2(F_\ell, F)}{\gamma_\ell} \leq C\right\}.\end{aligned}\quad (7.19)$$

Observe now that in view of (7.17) the inequality

$$\begin{aligned}\rho_2(Af_\ell, F) &\leq \sqrt{\gamma_\ell(C - W(f))} + \sqrt{\gamma_\ell(C - W(f)) + \gamma_\ell W(f)} \\ &= \sqrt{\gamma_\ell} \left(\sqrt{C - W(f)} + \sqrt{C} \right)\end{aligned}$$

is fulfilled in the domain

$$W(f) + \frac{\rho_2^2(F_\ell, F)}{\gamma_\ell} \leq C.$$

Since $\gamma_\ell \rightarrow 0$ as $\ell \rightarrow \infty$ for any δ starting with $\ell > n$, the inequality

$$P \left\{ \rho_2(Af_\ell, F) \leq \delta | W(f) + \frac{\rho_2^2(F_\ell, F)}{\gamma_\ell} \leq C \right\} = 1$$

is fulfilled. And since (7.19) is valid for all $\ell > n$, the inequality

$$P \left\{ \rho_1(f_\ell, f) \leq \varepsilon | W(f) + \frac{\rho_2^2(F_\ell, F)}{\gamma_\ell} \leq C \right\} = 1$$

is fulfilled. Thus it follows from (7.18) that for any $\delta > 0$ there exists n such that for all $\ell > n$ the inequality

$$P \{ \rho_1(f_\ell, f) \leq \varepsilon \} \geq P \left\{ W(f) + \frac{\rho_2^2(F_\ell, F)}{\gamma_\ell} \leq C \right\}$$

is fulfilled, and hence also the inequality

$$P \{ \rho_1(f_\ell, f) > \varepsilon \} \leq P \left\{ W(f) + \frac{\rho_2^2(F_\ell, F)}{\gamma_\ell} > C \right\}. \quad (7.20)$$

Taking into account that

$$C > W(f)$$

and introducing notation

$$\mu = C - W(f),$$

we obtain from (7.20) the assertion of the theorem:

$$P \{ \rho_1(f_\ell, f) > \varepsilon \} \leq P \{ \rho_2(F_\ell, F) > \sqrt{\mu \gamma_\ell} \}.$$

7.3.2 Proof of Theorem 7.2

1. An arbitrary closed sphere in Hilbert space (i.e., a set of vectors of the form $\{f: \|f - f_0\| \leq d\}$) is weakly compact. Therefore, as far as weak compactness in the space E_1 is concerned, we are under the conditions of Theorem 7.1. Consequently, for any positive ε and μ there exists a number $n = n(\varepsilon, \mu)$ such that for $\ell > n(\varepsilon, \mu)$ we obtain

$$P\{|(f_\ell, g) - (f, g)| > \varepsilon\} \leq P\{\rho_2(F_\ell, F) > \sqrt{\mu\gamma_\ell}\}, \quad (7.21)$$

where the expression

$$(q, f) = \int q(t)f(t) dt$$

defines the general form of linear functional in Hilbert space.

2. According to the definition of a norm in a Hilbert space we have

$$\begin{aligned} \|f_\ell - f\|^2 &= (f_\ell - f, f_\ell - f) \\ &= \|f_\ell\|^2 - \|f\|^2 + 2(f, f - f_\ell). \end{aligned}$$

Utilizing the inequality

$$P\{a + b > \varepsilon\} \leq P\left\{a > \frac{\varepsilon}{2}\right\} + P\left\{b > \frac{\varepsilon}{2}\right\},$$

we obtain

$$P\{\|f_\ell - f\|^2 > \varepsilon\} \leq P\left\{\|f_\ell\|^2 - \|f\|^2 > \frac{\varepsilon}{2}\right\} + P\left\{2(f, f - f_\ell) > \frac{\varepsilon}{2}\right\}. \quad (7.22)$$

In order to bound the first summand on the right-hand side we shall utilize the inequality (7.15), taking into account that

$$W(f) = \|f\|^2.$$

We thus obtain

$$\|f_\ell\|^2 \leq \|f\|^2 + \frac{\rho_2^2(F_\ell, F)}{\gamma_\ell}$$

Therefore

$$P\left\{\|f_\ell\|^2 - \|f\|^2 > \frac{\varepsilon}{2}\right\} \leq P\left\{\frac{\rho_2^2(F_\ell, F)}{\gamma_\ell} > \frac{\varepsilon}{2}\right\}$$

We bound the second summand (7.22) by means of (7.21), setting $\mu = \varepsilon/2$:

$$P \left\{ (f, f - f_\ell) > \frac{\varepsilon}{4} \right\} \leq P \left\{ \frac{\rho_2^2(F_\ell, F)}{\gamma_\ell} > \frac{\varepsilon}{2} \right\}.$$

Combining the last bound with (7.22), we arrive at the assertion of the theorem:

$$P \{ ||f_\ell - f||^2 > \varepsilon \} \leq 2P \left\{ \frac{\rho_2^2(F_\ell, F)}{\gamma_\ell} > \frac{\varepsilon}{2} \right\}.$$

The theorem thus has been proved.

7.3.3 Proof of Theorem 7.3

Since f_ℓ minimizes the functional

$$R_{\gamma_\ell}^*(\hat{f}, F_\ell, A_\ell) = \rho_2^2(A_\ell \hat{f}, F_\ell) + \gamma_\ell W(\hat{f}),$$

we have

$$\begin{aligned} \gamma_\ell W(f_\ell) &\leq R_{\gamma_\ell}^*(f_\ell, F_\ell, A_\ell) \leq R_{\gamma_\ell}^*(f, F_\ell, A_\ell) \\ &= \rho_2^2(A_\ell f, F_\ell) + \gamma_\ell W(f), \end{aligned} \quad (7.23)$$

where f is the desired solution of the equation. From (7.23) we find

$$W(f_\ell) \leq W(f) + \frac{\rho_2^2(A_\ell f, F_\ell)}{\gamma_\ell}$$

Since according to the triangle inequality we have

$$\begin{aligned} \rho_2(A_\ell f, F_\ell) &\leq \rho_2(A_\ell f, F) + \rho_2(F, F_\ell) \\ &\leq ||A_\ell - A|| W^{1/2}(f) + \rho_2(F, F_\ell), \end{aligned} \quad (7.24)$$

we obtain

$$W(f_\ell) \leq W(f) + \frac{1}{\gamma_\ell} \left(||A_\ell - A|| W^{1/2}(f) + \rho_2(F, F_\ell) \right)^2. \quad (7.25)$$

Taking into account that

$$\rho_2^2(A_\ell f_\ell, F_\ell) \leq R_{\gamma_\ell}^*(f_\ell, F_\ell, A_\ell)$$

from (7.25) and (7.23), we obtain

$$\rho_2^2(A_\ell f_\ell, F_\ell) \leq \gamma_\ell W(f) + \left(||A_\ell - A|| W^{1/2}(f) + \rho_2(F, F_\ell) \right)^2. \quad (7.26)$$

From this inequality using (7.24) and (7.12), we derive

$$\begin{aligned}
\rho_2(Af_\ell, F) &\leq \rho_2(Af_\ell, A_\ell f_\ell) + \rho_2(A_\ell f_\ell, F_\ell) + \rho_2(F_\ell, F) \\
&\leq W^{1/2}(f_\ell) \|A_\ell - A\| \\
&\quad + \left(\gamma_\ell W(f) + \left(\|A_\ell - A\| W^{1/2}(f) + \rho_2(F, F_\ell) \right)^2 \right)^{1/2} \\
&\quad + \rho_2(F_\ell, F) = \sqrt{\gamma_\ell} \left(\frac{\rho_2(F_\ell, F)}{\sqrt{\gamma_\ell}} + W^{1/2}(f_\ell) \frac{\|A_\ell - A\|}{\sqrt{\gamma_\ell}} \right) \\
&\quad + \sqrt{\gamma_\ell} \left(W(f) + \left(\frac{\rho_2(F_\ell, F)}{\sqrt{\gamma_\ell}} + W^{1/2}(f) \frac{\|A_\ell - A\|}{\sqrt{\gamma_\ell}} \right)^2 \right)^{1/2}.
\end{aligned} \tag{7.27}$$

Let us choose arbitrary constants $C_1, C_2 > 0$. Consider two events:

$$\mathcal{A} = \left\{ \omega : \frac{\rho_2(F, F_\ell)}{\sqrt{\gamma_\ell}} \leq C_1 \right\}, \quad \mathcal{B} = \left\{ \omega : \frac{\|A_\ell - A\|}{\sqrt{\gamma_\ell}} \leq C_2 \right\}.$$

Suppose that event \mathcal{A} and event \mathcal{B} occur simultaneously. Then from (7.25) we obtain

$$W(f_\ell) \leq W(f) + (C_1 + C_2 W^{1/2}(f))^2 = d < \infty. \tag{7.28}$$

From (7.28) and inequality (7.27) we obtain

$$\rho_2(Af_\ell, Af) \leq 2d\sqrt{\gamma_\ell}. \tag{7.29}$$

Note that according to the properties of the functional $W(f)$, inequality (7.28) means that the solution f_ℓ belongs to some compactum. According to the lemma about the inverse operator, considered in the Appendix to Chapter 1, the inverse operator A^{-1} is continuous on this compactum; that is, for any $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\rho_1(f, f_\ell) \leq \varepsilon \tag{7.30}$$

as soon as inequality

$$\rho_2(Af_\ell, Af) \leq \delta$$

is satisfied. According to (7.29) this inequality holds if

$$\gamma_\ell \leq \gamma_0 = \left(\frac{\delta}{2d} \right)^2.$$

Therefore for any $\varepsilon > 0$ there exists $\gamma_0 > 0$ such that for all $\gamma_\ell \leq \gamma_0$ events A and B imply the inequality (7.30). In other words, for all $\gamma_\ell \leq \gamma_0$ the inequality

$$\begin{aligned} P\{\rho_1(f_\ell, f) > \varepsilon\} &\leq P\{\bar{A} \cup \bar{B}\} \leq P\{\bar{A}\} + P\{\bar{B}\} \\ &= P\left\{\frac{\rho_2(F_\ell, F)}{\sqrt{\gamma_\ell}} > C_1\right\} + P\left\{\frac{\|A_\ell - A\|}{\sqrt{\gamma_\ell}} > C_2\right\} \end{aligned}$$

Here $\gamma_0 = \gamma(C_1, C_2, W(f), A^{-1}, \varepsilon)$; C_1, C_2, ε are arbitrary fixed constants. Theorem 7.3 has been proved.

7.4 CONDITIONS FOR CONSISTENCY OF THE METHODS OF DENSITY ESTIMATION

Consider the specific stochastic ill-posed problems: the problem of density estimation and the problem of the ratio of two densities estimation.

Consider the problem of density estimation from one of the following normed spaces:

- Hilbert space H with the norm

$$w_H(f) = \sqrt{\int_0^1 f^2(x) dx}.$$

- Sobolev space S_m with the norm

$$w_{S_m}(f) = \sqrt{\sum_{k=0}^m a_k \int_0^1 (f^{(k)}(x))^2 dx},$$

where $f^{(k)}(x)$ is the k th derivative of the function $f(x)$, and $a_k > 0$ are some constants.

- Space of smooth continuous on functions on $[0, 1]$ with bounded variation $V(f)$, possessing an m th derivative ($m \geq 0$) that satisfies the Lipschitz condition of order 6 ($6 > 0$):

$$\sup_{x,y \in [a,b]} \frac{|f^{(m)}(x) - f^{(m)}(y)|}{|x - y|^\delta} < \infty$$

(if $m > 0$, then function f has bounded variation $V(f)$; the requirement of bounded variation is essential when $m = 0$). For this space we consider the norm

$$w_{C_k}(f) = V(f) + \sum_{k=0}^m \sup_{x \in [0,1]} |f^{(k)}(x)| + \sup_{x,y \in [a,b]} \frac{|f^{(m)}(x) - f^{(m)}(y)|}{|x - y|^\delta}$$

In all these cases we consider the regularization functional $W(f)$ of the form

$$W(f) = w^2(f).$$

It is easy to check that functional $W(f)$ possesses all properties that are required for regularization terms.

Now let us consider the problem of estimating a density belonging to one of the sets of functions described above.

That is, we have to solve the integral equation

$$\int_0^x p(t) dt = F(x), \quad (7.31)$$

where instead of the right-hand side of Eq. (7.31) only the sequence of approximations

$$F_\ell(x) = \frac{1}{\ell} \sum_{i=1}^{\ell} \theta(x - x_i)$$

constructed on the basis of observations

$$x_1, \dots, x_\ell, \dots$$

is available.

For solving this equation we will use the regularization method; that is, we will minimize functional

$$R_{\gamma_\ell}(f) = \rho_{E_2}^2(Af, F_\ell) + \gamma_\ell W(f),$$

where $\rho_{E_2}(\cdot, \cdot)$ is the metric in space of (absolutely continuous) functions $F(x)$, and A is the operator

$$Af = \int_0^x f(t) dt.$$

In the space E_2 we will use one of the metrics:

1. Metric of the $L_1(0, 1)$ space:

$$\rho_{E_2}(F_1, F_2) = \int_0^1 |F_1(x) - F_2(x)| dx.$$

2. Metric of the $L_2(0, 1)$ space:

$$\rho_{E_2}(F_1, F_2) = \sqrt{\int_0^1 (F_1(x) - F_2(x))^2 dx}.$$

3. Metric of the $C(0, 1)$ space:

$$\rho_{E_2}(F_1, F_2) = \sup_x |F_1(x) - F_2(x)|$$

Since the distance in metric C is not less than in metrics $L_1(0, 1)$ or $L_2(0, 1)$, the bound obtained with this metric will be also valid for the other two metrics.

Suppose that

$$f_1, \dots, f_\ell, \dots$$

is a sequence of the solutions obtained by means of the regularization method. Then according to Theorem 7.1 for any ε and any μ the inequality

$$P\{\rho_{E_1}(f_\ell, f) > \varepsilon\} \leq P\{\sup_x |F_\ell(x) - F(x)| > \sqrt{\gamma_\ell \mu}\}$$

holds true for sufficiently large C . According to the Kolmogorov inequality (see Chapter 1, Section 11.3) for sufficiently large ℓ , one has

$$P\{\sup_x |F_\ell(x) - F(x)| > \varepsilon\} \leq 2 \exp\{-2\varepsilon^2\ell\}.$$

Therefore there exists a $\ell(\varepsilon, \mu)$ such that for $\ell > \ell(\varepsilon, \mu)$ the inequality

$$P\{\rho_{E_1}(f_\ell, f) > \varepsilon\} \leq 2 \exp\{-2\gamma_\ell \mu \ell\} \quad (7.32)$$

is fulfilled.

If $f(x) \in L_2(0, 1)$, it then follows from Theorem 7.2 and the Kolmogorov inequality that for $\ell > \ell(\varepsilon)$, the inequality

$$P\left\{\int (f_\ell(x) - f(x))^2 dx > \varepsilon\right\} \leq 2 \exp\{-2\gamma_\ell \mu \ell\} \quad (7.33)$$

holds. Inequalities (7.32) and (7.33) imply that the solution f_ℓ converges in probability to the desired one (in the metric $\rho_{E_1}(f_\ell, f)$) if

$$\begin{aligned} \gamma_\ell &\xrightarrow[\ell \rightarrow \infty]{} 0, \\ \ell \gamma_\ell &\xrightarrow[\ell \rightarrow \infty]{} \infty. \end{aligned} \quad (7.34)$$

(In this case the right-hand sides of Eqs. (7.32) and (7.33) converge to zero.)

Inequalities (7.32) and (7.33) also imply that the solution f_ℓ converges almost surely to the desired one (in the metric $\rho_{E_1}(f_\ell, f)$) if

$$\begin{aligned} \gamma_\ell &\xrightarrow[\ell \rightarrow \infty]{} 0, \\ \gamma_\ell \frac{\ell}{\ln \ell} &\xrightarrow[\ell \rightarrow \infty]{} \infty. \end{aligned} \tag{7.35}$$

(In this case for all $\varepsilon > 0$ and $\mu > 0$ the conditions of Borel–Cantelli lemma

$$\begin{aligned} &\sum_{\ell=1}^{\infty} P\{\rho_{E_1}(f_\ell, f) > \varepsilon\} \\ &< \sum_{\ell=1}^{\infty} P\{\sup_x |F_\ell(x) - F(x)| > \sqrt{\gamma_\ell \mu}\} < 2 \sum_{\ell=1}^{\infty} e^{-2\gamma_\ell \mu \ell} < \infty \end{aligned}$$

hold true.)

7.5 NONPARAMETRIC ESTIMATORS OF DENSITY: ESTIMATORS BASED ON APPROXIMATIONS OF THE DISTRIBUTION FUNCTION BY AN EMPIRICAL DISTRIBUTION FUNCTION

This section shows that by using the regularization method with empirical distribution functions $F_\ell(x)$ instead of the unknown function $F(x)$, one can obtain the classical nonparametric density estimators: Parzen's windows, projective estimators, spline estimators, and so on.

Note, however, that approximations $F_\ell(x)$ do not reflect an important property of the distribution functions for which there exist densities. These distribution functions $F(x)$ belong to the set of absolutely continuous *functions*, while the approximations $F_\ell(x)$ are discontinuous functions. As we will see in the next section, the continuous (spline) approximations to an unknown distribution function implies new nonparametric estimators that differ from the classical ones.

7.5.1 The Parzen Estimators

Estimators for Unbounded Support Let us specify the functional

$$R(f, F_\ell) = \rho_2^2(Af, F_\ell) + \gamma_\ell W(f). \tag{7.36}$$

Below we use:

1. Distance in the L_2 metric:

$$\rho_{E_2}(F, F_\ell) = \sqrt{\int_{-\infty}^{\infty} (F(x) - F_\ell(x))^2 dx}.$$

2. Regularization functional of the form:

$$W(f) = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} K(z-x)f(x)dx \right)^2 dz$$

Here $K(z-x)$ is the kernel of the linear operator

$$Bf = \int_{-\infty}^{\infty} K(z-x)f(x)dx.$$

In particular if $K(z-x) = \delta^p(z-x)$ the operator B defines the p th derivative of the function $f(x)$.

For these elements we have the functional

$$\begin{aligned} R_{\gamma_\ell}(f, F_\ell) \\ = \int_{-\infty}^{\infty} \left(\int_{-\infty}^x f(t)dt - F_\ell(x) \right)^2 dx + \gamma_\ell \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} K(z-x)f(x)dx \right)^2 dz. \end{aligned} \quad (7.37)$$

We show that estimators f_γ that minimize this functional are Parzen's windows. Indeed, let us denote by $\bar{f}(\omega)$ the Fourier transform of the function $f(t)$ and by $\bar{K}(\omega)$ the Fourier transform of the function $K(x)$. Then one can evaluate the Fourier transform for the functions

$$\begin{aligned} \bar{F}(\omega) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} F(x)e^{-i\omega x} dx \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-i\omega x} \int_{-\infty}^x f(t)dt = \frac{\bar{f}(\omega)}{i\omega} \\ \bar{F}_\ell(\omega) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} F_\ell(x)e^{-i\omega x} dx = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{1}{\ell} \sum_{j=1}^{\ell} \theta(x-x_j)e^{-i\omega x} dx \\ &= \frac{1}{\ell} \sum_{j=1}^{\ell} \frac{e^{-i\omega x_j}}{i\omega}. \end{aligned}$$

Note that the Fourier transform for convolution of two functions is equal to the product of the Fourier transforms of these two functions. For our case this means that

$$\begin{aligned} \frac{1}{2\pi} \int_{-\infty}^{\infty} (K(x)*f(x))e^{-i\omega x} dx &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} K(z-x)f(x)dx \right) e^{-i\omega z} dz \\ &= \bar{K}(\omega)\bar{f}(\omega). \end{aligned}$$

Lastly, recall that according to Parseval's equality the L_2 norm of any function $f(x)$ within the constant $1/2\pi$ is equal to the L_2 norm of its Fourier transform

$f(\omega)$ (here $\tilde{f}(\omega)$ is the Fourier transform of the function $f(x)$). Therefore one can rewrite (7.37) in the form

$$R_{\gamma_\ell}(f, F_\ell) = \left\| \frac{\tilde{f}(\omega) - \frac{1}{\ell} \sum_{j=1}^{\ell} e^{-i\omega x_j}}{i\omega} \right\|_{L_2}^2 + \gamma_\ell \|\bar{K}(\omega) \tilde{f}_\ell(\omega)\|_{L_2}^2$$

This functional is quadratic with respect to $\tilde{f}(\omega)$.

Therefore the condition for its minimum is

$$\frac{\tilde{f}_\ell(\omega)}{\omega^2} - \frac{1}{\ell \omega^2} \sum_{j=1}^{\ell} e^{i\omega x_j} + \gamma_\ell \bar{K}(\omega) K(-\omega) f(w) = 0. \quad (7.38)$$

Solving this equation with respect to $\tilde{f}_\ell(\omega)$, one obtains

$$\tilde{f}_\ell(\omega) = \left(\frac{1}{1 + \gamma_\ell \omega^2 \bar{K}(\omega) \bar{K}(-\omega)} \right) \frac{1}{\ell} \sum_{j=1}^{\ell} e^{-i\omega x_j}.$$

Let us introduce the notations

$$g_{\gamma_\ell}(\omega) = \frac{1}{1 + \gamma_\ell \omega^2 \bar{K}(\omega) \bar{K}(-\omega)}$$

and

$$G_{\gamma_\ell}(x) = \int_{-\infty}^{\infty} g_{\gamma_\ell}(\omega) e^{i\omega x} d\omega.$$

To obtain an approximation to the density, one has to evaluate the inverse Fourier transform

$$\begin{aligned} f_\ell(x) &= \int_{-\infty}^{\infty} \tilde{f}_\ell(\omega) e^{i\omega x} d\omega = \int_{-\infty}^{\infty} g_{\gamma_\ell}(\omega) \left(\frac{1}{\ell} \sum_{j=1}^{\ell} e^{-i\omega x_j} \right) e^{i\omega x} d\omega \\ &= \frac{1}{\ell} \sum_{j=1}^{\ell} \int_{-\infty}^{\infty} g_{\gamma_\ell}(\omega) e^{i\omega(x-x_j)} d\omega = \frac{1}{\ell} \sum_{j=1}^{\ell} G_{\gamma_\ell}(x - x_j). \end{aligned}$$

The last expression is the Parzen estimator with kernel $G_{\gamma_\ell}(x)$. Using different regularizer functions (different functions $K(u)$ in (7.37))' one can obtain different Parzen's kernels.

Let us consider the important case

$$K(x) = \delta^{(p+1)}(x),$$

where $\delta(x)$ is the Dirac delta function, $p \geq 0$ (that is, we consider the case where the desired function possesses p derivatives).

For this case the kernel is

$$G_{\gamma,p}(x) = \int_{-\infty}^{\infty} \frac{e^{ix\omega}}{1 + \gamma \omega^{2(p+1)}} d\omega.$$

After integration, one obtains

$$G_{\gamma,p}(x) = \frac{1}{2(p+1)\lambda(p)} \sum_{r=0}^p \sin \left(a_r + \frac{x}{\lambda(p)} \cos a_r \right) e^{-|x|\sin a_r/\lambda(p)}, \quad (7.39)$$

where we denote

$$\lambda(p) = \gamma^{1/2(p+1)}, \quad a_r = \frac{\pi(1+2r)}{2(p+1)}, \quad r = 0, \dots, p.$$

For the case $p = 0$ (the desired density belongs to L_2), Parzen's kernel is

$$G_{\gamma,0}(x) = \frac{1}{2\sqrt{\gamma}} \exp \left\{ -\frac{|x|}{\sqrt{\gamma}} \right\}.$$

For the case $p = 1$ (the derivative of the desired density belongs to L_2), Parzen's kernel is

$$G_{\gamma,1}(x) = \frac{1}{2\sqrt[4]{4\gamma}} \exp \left\{ -\frac{|x|}{\sqrt[4]{4\gamma}} \right\} \left(\cos \frac{x}{\sqrt[4]{4\gamma}} + \sin \frac{|x|}{\sqrt[4]{4\gamma}} \right).$$

Estimators for Bounded Support. The estimators considered above were derived under the condition that the support of the density is infinite (to derive Parzen's estimator we used Fourier transform). If the density concentrates on the finite support $a < x < b$, then Parzen's window gives a bias estimate of the density. Indeed, for this case the integral is less than one:

$$\int_a^b \bar{f}(x) dx = \frac{1}{\ell} \sum_{i=1}^{\ell} \int_a^b G_{\gamma}(x) dx < 1$$

(it equals one on the infinite support). Below, for this case we will derive another estimator. Using the same method for deriving an estimator, one can obtain corrections to Parzen's window, which make it suitable for a case with finite support.

For simplicity we shall stipulate in relation to (7.37)

$$\begin{aligned} \rho_{E_2}(F, F_{\ell}) &= \sqrt{\int_a^b (F(x) - F_{\ell}(x))^2 dx}, \\ W(f) &= \int_a^b f^2(t) dt. \end{aligned}$$

We then obtain the functional

$$R(f, F_\ell) = \int_a^b \left(\int_a^x f(t) dt - F_\ell(x) \right)^2 dx + \gamma_\ell \int_a^b f^2(t) dt,$$

which we rewrite in the form

$$R(f, F_\ell) = \int_a^b \left[(F(x) - F_\ell(x))^2 + \gamma_\ell \dot{F}^2(x) \right] dx,$$

where we denote by $\dot{F}(x)$ the first derivative of the function $F(x)$. The minimum of this functional (in the set of functions $F(x)$ for which the first derivative has a finite norm in $L_2(a, b)$) satisfies the Euler equation

$$\begin{aligned} \gamma_\ell \dot{F}(x) - (F(x) - F_\ell(x)) &= 0 \\ F(a) = 0, \quad F(b) = 1. \end{aligned} \tag{7.40}$$

Let $F_{\gamma_\ell}(x)$ be the solution of Eq. (7.40). Then the estimate of the density is

$$f_\ell(x) = \dot{F}(x).$$

The solution of the linear equation (which is a sum of the particular solution and the general solution) is

$$\begin{aligned} f_\gamma(t) &= \frac{1}{2\ell\sqrt{\gamma_\ell}} \sum_{i=1}^\ell \exp \left\{ -\frac{|x - x_i|}{\sqrt{\gamma_\ell}} \right\} \\ &\quad + C_1 \exp \left\{ -\frac{|x - a|}{\sqrt{\gamma_\ell}} \right\} + C_2 \exp \left\{ -\frac{|x - b|}{\sqrt{\gamma_\ell}} \right\}, \end{aligned} \tag{7.41}$$

where constants C_1 and C_2 are determined by conditions $F(a) = 0$ and $F(b) = 1$ are given by

$$\begin{aligned} C_1 &= \frac{1}{2\ell\sqrt{\gamma_\ell}} \sum_{i=1}^\ell \frac{\exp \left\{ \frac{b - x_i}{\sqrt{\gamma_\ell}} \right\} + \exp \left\{ \frac{x_i - b}{\sqrt{\gamma_\ell}} \right\}}{\exp \left\{ \frac{b - a}{\sqrt{\gamma_\ell}} \right\} - \exp \left\{ \frac{a - b}{\sqrt{\gamma_\ell}} \right\}}, \\ C_2 &= \frac{1}{2\ell\sqrt{\gamma_\ell}} \sum_{i=1}^\ell \frac{\exp \left\{ \frac{a - x_i}{\sqrt{\gamma_\ell}} \right\} + \exp \left\{ \frac{x_i - a}{\sqrt{\gamma_\ell}} \right\}}{\exp \left\{ \frac{b - a}{\sqrt{\gamma_\ell}} \right\} - \exp \left\{ \frac{a - b}{\sqrt{\gamma_\ell}} \right\}}. \end{aligned}$$

Therefore the estimate for density in the Parzen form has a regular Parzen estimate and corrections for the ending points of the finite support. It is easy to see that for infinite support the constants C_1 and C_2 are equal to zero.

7.5.2 Projection Estimators

Let the desired density $f(x)$ belong to the set of functions whose p th derivative ($p \geq 0$) belongs to $L_2(0, \pi)$. We looking for an estimator of the density, $f_\ell(x)$, that minimizes the functional

$$R_{\gamma_\ell}(f, F_\ell) = \int_0^\pi \left(\int_0^x f(t) dt - F_\ell(x) \right)^2 dx + \gamma_\ell \int_0^\pi (f^{(p)}(t))^2 dt. \quad (7.42)$$

Consider the approximation to the unknown density given by an expansion in the orthonormal functions $\phi_1(t), \dots, \phi_k(t), \dots$. We consider the expansion in $\cos rt$, $n = 0, 1, \dots, k, \dots$, which gives the simplest estimator.

Let us consider the approximation to the desired function in the form

$$f_\gamma(t) = \frac{1}{2} + \sum_{r=1}^{\infty} a_r \cos rt, \quad (7.43)$$

where a_1, \dots, a_r, \dots are coefficients. Putting this approximation into (7.42), one obtains that the minimum of this functional is reached when

$$a_r = \frac{\frac{1}{\ell} \sum_{i=1}^{\ell} \cos rt_i}{1 + \gamma_\ell r^{2(p+1)}}, \quad r = 1, 2, \dots$$

Note that two estimators—estimator (7.39) with $a = 0$ and $b = \pi$ and estimator (7.43) with $p = 0$ —describe on $(0, \pi)$ the same function since both estimators are the functions that minimize the same functional.

7.5.3 Spline Estimate of the Density. Approximation by Splines of the Odd Order

Let the p th derivative of the density be a square integrable in (a, b) . Let us consider the functional which is constructed by the metric $L_1(a, b)$

$$\rho_{E_2}(F, F_\ell) = \int_a^b |F(x) - F_\ell(x)| dx$$

and regularizer

$$W(f) = \int_a^b (f^{(p)}(t))^2 dt$$

Therefore we have to minimize the functional

$$R_{\gamma_\ell}(F, F_\ell) = \left(\int_a^b |F(x) - F_\ell(x)| dx \right)^2 + \gamma_\ell \int_a^b (f^{(p)}(t))^2 dt. \quad (7.44)$$

Consider the functional which slightly differs from (7.44):

$$R_{\gamma_\ell}^*(F, F_\ell) = \int_a^b |F(x) - F_\ell(x)| dx + \mu(\gamma_\ell) \int_a^b (f^{(p)}(t))^2 dt.$$

Let us rewrite this functional in the form that is used in optimal control theory:

$$\begin{aligned} z_0(t) &= F(t), \\ z_i(t) &= f^{(i-1)}(t), \quad i = 1, 2, \dots, (p-1), \\ f^{(p)}(t) &= u(t). \end{aligned}$$

In this notation the problem of control theory is to minimize the functional

$$R^* = \int_a^b |z(x) - F_\ell(x)| dx + \mu_\ell \int u^2(x) dx$$

under restrictions

$$\begin{aligned} \dot{z}_i(x) &= z_{i+1}(x), \quad i = 1, 2, \dots, (p-1), \\ \dot{z}_p(x) &= u(x). \end{aligned}$$

The solution of this optimal control problem is the spline of odd order $2p+1$

$$f^{(2p+1)}(x) = \frac{(-1)^p}{\mu_\ell} \text{sign}(F(x) - F_\ell(x))$$

satisfying the conditions

$$f^r(a) = f^r(b) = 0, \quad r = p+1, \dots, 2p+1$$

For $p \geq 1$ the approximations converge to the desired solution in the $C^{p-1}(a, b)$ metric, and for $p = 0$ the approximations converge to the desired solution in the $L_2(a, b)$ metric.

7.5.4 Spline Estimate of the Density. Approximation by Splines of the Even Order

As in the previous case, let the p th derivative of the desired density be a function that has a finite $L_2(a, b)$ norm. Consider the regularizing functional

$$W(f) = \int_a^b (f^{(p)}(x))^2 dx$$

and the $C(a, b)$ metric

$$\rho_{E_2}(F, F_\ell) = \sup_{a \leq x \leq b} |F(x) - F_\ell(x)|$$

Therefore we are looking for an estimator that minimizes the functional

$$R(f, F_\ell) = \left(\max_{a \leq x \leq b} |F(x) - F_\ell(x)| \right)^2 + \gamma_\ell \int_a^b \left(f^{(p)}(x) \right)^2 dx.$$

One can prove (Aidu and Vapnik, 1989) that the solution to this optimization problem is a spline of order $2p$ (for $p = 0$ it is the histogram) that possesses the following property: The interval $[a, b]$ is divided into subintervals such that the maximal difference $F_\ell(x)$ and $F(x)$ in each subinterval takes the same absolute values and alternates in sign from interval to interval.

7.6 NONCLASSICAL ESTIMATORS

7.6.1 Estimators for the Distribution Function

In the last section we derived the main classical nonparametric estimators using the regularization method for solving ill-posed problems. Along with this general principle for constructing the estimators, common to all these examples was the use of the empirical distribution function as an approximation to the unknown distribution function

$$F_\ell(x) = \frac{1}{\ell} \sum_{i=1}^{\ell} \theta(x - x_i). \quad (7.45)$$

As mentioned previously, the approximation (7.45) does not include the whole a priori information about the unknown distribution function. It is known that any distribution function that has a density is absolutely continuous; however, our approximations are discontinuous functions. Note also that we use the approximation for solving ill-posed problems, where small changes on the right-hand side can cause large changes in the solution. In Section 7.4 we proved that the densities obtained on the basis of the approximations $F_\ell(x)$ converge to the desired one as well as densities obtained on the basis of any other approximation, satisfying Kolmogorov's inequality. However, for a finite number of observations the obtained solutions can be very different.

Therefore the following problem arises: How do we construct approximations to the unknown distribution function which converge as fast as the empirical distribution function $F_\ell(x)$ and satisfy the additional property to be a continuous monotonic function?

Below we will construct these approximations and show that they imply new estimators for density.

7.6.2 Polygon Approximation of Distribution Function

Let us consider the polygonal approximation to a one-dimensional distribution function. Let

$$\bar{x}_1, \dots, \bar{x}_\ell$$

be an ordered array of the sample x_1, \dots, x_ℓ . Consider the following approximation to the distribution function, called the *polygon approximation*:

$$F_\ell^{\text{pol}}(x) = \begin{cases} 0 & \text{if } x < \bar{x}_1, \\ \frac{1}{\ell} & \text{if } \bar{x}_1 \leq x < \frac{\bar{x}_1 + \bar{x}_2}{2}, \\ \frac{k}{\ell} + \frac{1}{\ell} \frac{2x - \bar{x}_k + \bar{x}_{k+1}}{\bar{x}_{k+1} - \bar{x}_{k-1}} & \text{if } \frac{\bar{x}_{k-1} + \bar{x}_k}{2} \leq x < \frac{\bar{x}_k + \bar{x}_{k+1}}{2}, \quad k < \ell - 1, \\ 1 - \frac{1}{\ell} & \text{if } \frac{\bar{x}_{\ell-1} + \bar{x}_\ell}{2} \leq x < \bar{x}_\ell, \\ 1 & \text{if } x \geq \bar{x}_\ell. \end{cases} \quad (7.46)$$

In Fig. 7.1 the approximations to the distribution functions are presented: Figure 7.1a shows the empirical distribution function and Fig 7.1b shows the polygonal approximation to the distribution function.[†]

7.6.3 Kernel Density Estimator

Consider the functional (7.37) whose point of minimum forms the Parzen's estimator under the condition that one uses the empirical distribution function $F_\ell(x)$.

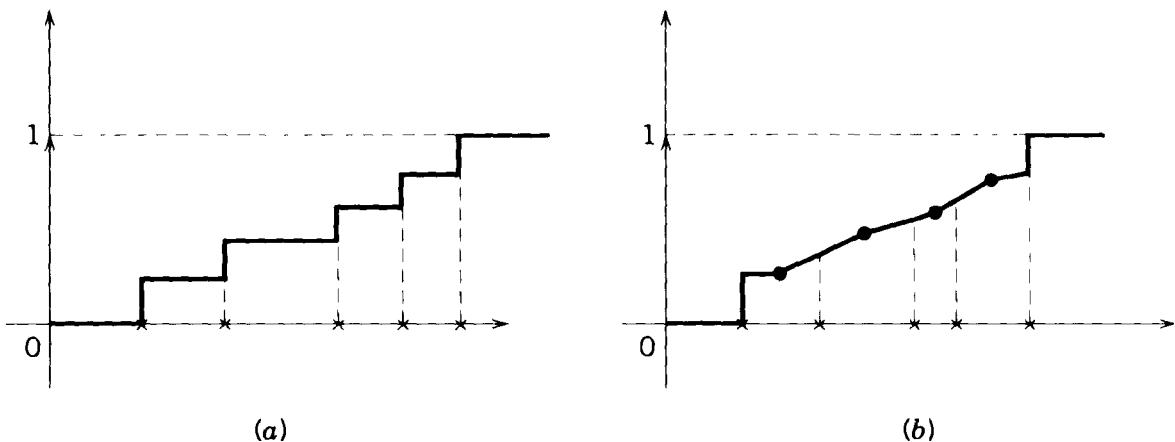


FIGURE 7.1. Empirical distribution function (a) and polygon distribution function (b). Note that approximation (b) has two discontinuous points.

[†] More generally, one can use a spline approximation of order $d \geq 0$.

Now in this functional we use the polygonal approximation (7.45) instead of the empirical distribution function F_ℓ :

$$\begin{aligned} R_\gamma(f, F_\ell^{\text{pol}}) \\ = \int_{-\infty}^{\infty} \left(\int_{-\infty}^x f(t) dt - F_\ell^{\text{pol}}(x) \right)^2 dx + \gamma_\ell \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} K(z-x)f(x) dx \right)^2 dz. \end{aligned}$$

To obtain the estimator we use the Fourier transform representation that is valid for an arbitrary approximation to a desired density function

$$\bar{f}(\omega) = \frac{1}{2\pi} \bar{g}(\omega) \bar{f}_\ell(\omega),$$

where

$$\bar{g}_\gamma(\omega) = \frac{1}{1 + \gamma_\ell \omega^2 \bar{K}(\omega) \bar{K}(-\omega)}$$

is the Fourier transform of the kernel in the Parzen's estimator, and $\bar{f}_\ell(\omega)$ is the Fourier transform of the empirical density estimator $\dot{F}_\ell(x) = f_\ell(x)$.

Note that for the Fourier transform the equality

$$f(x) = \int_{-\infty}^{\infty} \bar{g}_\gamma(\omega) \bar{f}_\ell(\omega) e^{ix\omega} d\omega = \int_{-\infty}^{\infty} g(x-z) f_\ell(z) dz \quad (7.47)$$

holds true, where $g(u)$ is the kernel in a Parzen estimator.

In our case for the polygon approximation of the distribution function we have

$$\begin{aligned} f_\ell(x) \\ = \frac{1}{\ell} \left(\delta(x - \bar{x}_1) + \delta(x - \bar{x}_\ell) + \sum_{i=2}^{\ell-1} 2 \frac{\theta(x - \frac{\bar{x}_{i+1} + \bar{x}_i}{2}) - \theta(x - \frac{\bar{x}_i + \bar{x}_{i-1}}{2})}{\ell(\bar{x}_{i+1} - \bar{x}_{i-1})} \right). \end{aligned}$$

Putting $f_\ell(x)$ in (7.47) we obtain the estimator

$$\begin{aligned} f(x) \\ = \frac{1}{\ell} \left(g_\gamma(x - \bar{x}_1) + g_\gamma(x - \bar{x}_\ell) + \sum_{i=1}^{\ell-1} \frac{2}{\bar{x}_{i+1} - \bar{x}_{i-1}} \int_{\frac{\bar{x}_{i+1} + \bar{x}_i}{2}}^{\frac{\bar{x}_{i+1} + \bar{x}_i}{2}} g_\gamma(z - x) dz \right). \end{aligned}$$

In contrast to Parzen's kernel estimator, this type of kernel estimator depends on the distance between two neighboring points of the ordered array row of

sample. Both types of kernel estimators (Parzen's and the new one) depend on the value of the regularization parameter γ_ℓ . However, the new estimator has two characteristics of width: local, which depends on the distance between two elements of the ordered array; and global, which depends on the regularization parameter γ_ℓ (see Fig. 7.2).

7.6.4 Projection Method of the Density Estimation

Consider the functional (7.42), the minimum of which defines the projection methods for density estimation under the condition that one uses the empirical distribution function. Now in this functional we use polygonal approximation (7.46) (instead of empirical distribution function $F_\ell(x)$):

$$R_{\gamma_\ell}, F_\ell = \left(\int_0^x f(t) dt - F_\ell^{\text{pol}}(x) \right)^2 dx + \gamma_\ell \int_0^\pi \left(f^{(p)}(t) \right)^2 dt. \quad (7.48)$$

As before, one considers the approximation to the unknown density as an expansion on the orthogonal on $(0, \pi)$ functions $\cos nt$, $n = 1, \dots$. We obtain

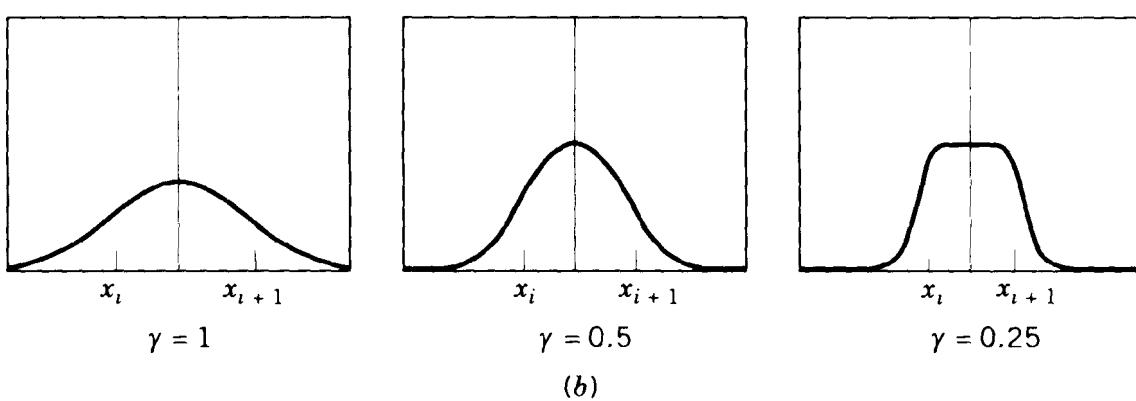
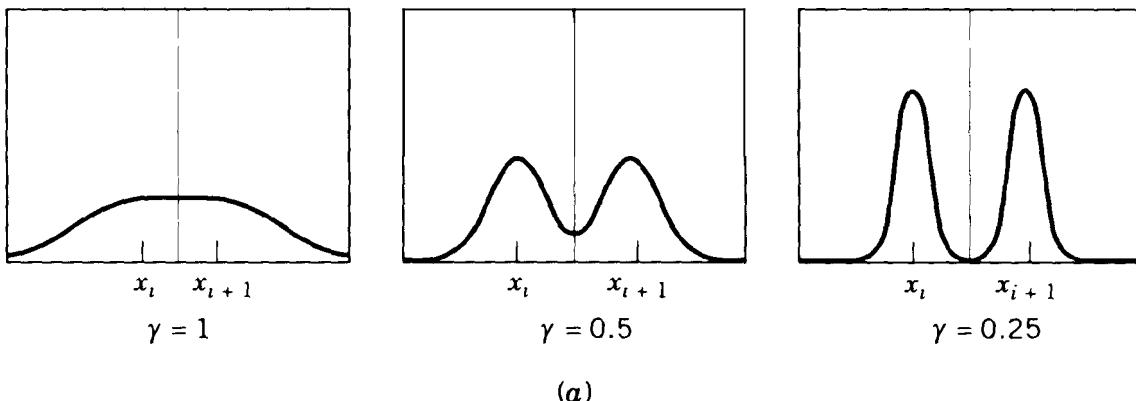


FIGURE 7.2. Kernels for different values of γ constructed using Gaussians. Parzen kernels (a) and new kernels (b).

the solution

$$f(t) = \frac{1}{\pi} + \frac{2}{\pi} \sum_{n=1}^{\infty} a_n \cos nt,$$

where

$$a_n =$$

$$\frac{1}{(1 + \gamma_\ell n^{2(k+1)})\ell} \left(\cos n\bar{x}_1 + \cos n\bar{x}_\ell + 2 \sum_{i=2}^{\ell-1} \frac{\sin n\frac{\bar{x}_{i+1} + \bar{x}_i}{2} - \sin n\frac{\bar{x}_i + x_{i-1}}{2}}{n(\bar{x}_{i+1} - \bar{x}_{i-1})} \right).$$

We restrict our discussion to these two examples of nonclassical density estimators. However, using continuous approximations of distribution functions in appropriate functionals, one can obtain new nonclassical estimators.

7.7 ASYMPTOTIC RATE OF CONVERGENCE FOR SMOOTH DENSITY FUNCTIONS

Section 7.4 proved the consistency of the regularization method under very weak conditions with regard to the rules for choosing the regularization parameter γ_ℓ (see (7.34) and (7.35)). In this section we consider a specific law for choosing γ_ℓ :

$$\gamma_\ell = \frac{\ln \ln \ell}{\ell}.$$

For this law we evaluate the asymptotic rate of convergence of the estimators to the unknown smooth density.

We shall apply the regularization method to estimate smooth densities defined on the interval $[a, b]$.

Suppose that it is known that the probability density $f(t)$ possesses m derivatives (m may be equal to zero), and let the function $f^{(m)}(t)$ satisfy the Lipschitz conditions of order μ ($0 \leq \mu \leq 1$):

$$|f^{(m)}(t) - f^{(m)}(\tau)| < K(f)|t - \tau|^\mu,$$

$$K(f) = \sup_{t, \tau \in [a, b]} \frac{|f^{(m)}(t) - f^{(m)}(\tau)|}{|t - \tau|^\mu}$$

Consider the following regularizer:

$$W(f) = \left(\max_{0 \leq k \leq m} \sup_{t \in [a, b]} |f^{(k)}(t)| + \sup_{t, \tau \in [a, b]} \frac{|f^{(m)}(t) - f^{(m)}(\tau)|}{|t - \tau|^\mu} \right)^2 \quad (7.49)$$

This functional is lower semicontinuous, and the set of functions

$$\mathcal{M}_c = \{f : W(f) \leq c\}$$

is compact in C . Therefore this functional can be used as a regularizer.

We consider $C[a, b]$ distance between the distribution function and its approximation

$$\rho_C(F, F_\ell) = \sup_{x \in [a, b]} |F(x) - F_\ell(x)|.$$

Therefore we consider the sequence of functions minimizing the functional

$$\begin{aligned} R_\ell(f, F_\ell) &= \left(\left| \sup_{x \in [a, b]} \int_a^b \theta(x-t)f(t) dt - F_\ell(x) \right| \right)^2 \\ &+ \frac{\ln \ln \ell}{\ell} \left(\max_{0 \leq k \leq m} \sup_{t \in [a, b]} |f^{(k)}(t)| + \sup_{t, \tau \in [a, b]} \frac{|f^{(m)}(t) - f^{(m)}(\tau)|}{|t - \tau|^\mu} \right)^2. \end{aligned} \quad (7.50)$$

In this section we shall estimate the asymptotic rate of convergence in the metric C of the sequence of solutions $f_\ell(t)$ to the required density. As will be shown below, the rate of convergence depends on the degree of smoothness of the estimated density, characterized by the quantity

$$\beta = m + \mu$$

(the larger β , the larger the rate).

Theorem 7.4. *An asymptotic rate of convergence of the approximations $f_\ell(t)$ to the required density $f(t)$ is determined by the expression*

$$P \left\{ \limsup_{\ell \rightarrow \infty} \left(\frac{\ln \ln \ell}{\ell} \right)^{\beta/(2\beta+2)} \sup_{a \leq t \leq b} |f(t) - f_\ell(t)| \leq g \right\} = 1,$$

where g is any constant.

In other words, the sequence of functions minimizing the risk functional (7.50) converges to the desired density function with the rate

$$r_\ell = \left(\frac{\ln \ln \ell}{\ell} \right)^{\beta/2(\beta+1)}.$$

Observe that the value of regularization parameter

$$\gamma_\ell = \frac{\ln \ln \ell}{\ell}$$

does not satisfy conditions (7.35). Nevertheless, it implies uniform convergence approximations $f_\ell(t)$ to the unknown density with probability one. The conditions (7.35) are only sufficient conditions.

Finally, the following should be mentioned before proceeding to the proof of the theorem. In 1978 Hasminskii obtained an estimate for the best possible rate of convergence of an approximation to an unknown density. He discovered that there is no algorithm which would ensure convergence in $C[a, b]$ to a β -smooth density at a rate whose order of magnitude is larger than

$$r_\ell = \left(\frac{\ln \ell}{\ell} \right)^{\beta/(2\beta+1)}.$$

It was shown that using Parzen's window, one can achieve this rate of convergence.

This result is slightly better than the result obtained in Theorem 7.4. However, one has to note that the maximal rate of convergence was achieved for a special kernel ($K(u) = \sin u/u$) rather than for kernels that are easy to handle and usually used.

For the regularization method the same best rate of convergence can be obtained if one uses the special metric $\rho(F, F_\ell)$ in the functional (7.50).

To construct this metric, divide the interval $[a, b]$ into

$$n = \left(\frac{\ell}{\ln \ell} \right)^{1/(2\beta+1)}$$

equal parts

$$[x_i, x_{i+1}), \quad x_i = a + i \frac{b-a}{n}, \quad i = 1, \dots, n,$$

and define the quantities

$$\|F(x) - F_\ell(x)\|_i = |F(x_{i+1}) - F_\ell(x_{i+1}) - F(x_i) + F_\ell(x_i)|.$$

Using these quantities, one can construct the functional

$$R_\ell^*(f, F_\ell) = \left(\sup_{1 \leq i \leq n} \left\| \int_a^b \theta(x-t)f(t) dt - F_\ell(x) \right\|_i \right)^2 + \frac{\ln \ell}{\ell} W(f),$$

where $W(f)$ is given in (7.49).

The theorem is valid (its proof is analogous to that of Theorem 7.4), which asserts that the sequence of functions minimizing $R_\ell^*(f, F_\ell)$ converges as ℓ increases in $C[a, b]$ to a β -smooth density $f(t)$ at a rate whose order of magnitude is the best obtainable:

$$P \left\{ \limsup_{\ell \rightarrow \infty} \left(\frac{\ell}{\ln \ell} \right)^{\beta/(2\beta+1)} \sup_{a \leq t \leq b} |f(t) - f_\ell(t)| \leq g \right\} = 1.$$

7.8 PROOF OF THEOREM 7.4

To prove the theorem, the following lemma will be required.

Lemma 7.1. Consider a function $y(t)$ that is $m+1$ times continuously differentiable on the interval $[a,b]$. Denote by $x(t)$ the derivative of this function. Let the m th ($m \geq 0$) derivative of $x(t)$ satisfy the Lipschitz condition of the order μ on $[a,b]$:

$$\sup_{t,\tau \in [a,b]} |x^m(t) - x^m(\tau)| \leq K|t - \tau|^\mu.$$

Then the inequality

$$\|x\|_C \leq \max \left\{ C_m^* \|y\|_C; C_m^{**} \|y\|_C^{(m+\mu)/(1+m+\mu)} \right\}$$

is valid, where

$$\begin{aligned} \|z\|_C &= \sup_{t \in [a,b]} |z(t)|, \\ C_m^* &= \frac{2^{(1+2m)}}{b-a} \left(\frac{1+\mu}{\mu} \right), \\ C_m^{**} &= \left[2^{(m+\mu)(1+m+\mu)-\mu^2} K \left(\frac{\mu+1}{\mu} \right)^\mu \right]^{1/(1+m+\mu)} \end{aligned}$$

Proof

1. Consider first the case of $m = 0$. Choose on $[a,b]$ an arbitrary point t^* such that $|x(t^*)| \neq 0$. Define an ε -neighborhood of this point with

$$\varepsilon = \left(\frac{|x(t^*)|}{K} \right)^{1/\mu} \quad (7.51)$$

Assume that at least one of the endpoints of this neighborhood—say the right one—is located within the interval $[a,b]$; that is, $t^* + \varepsilon \leq b$. Along with the function $x(t)$ consider the function

$$\phi(\tau) = |x(t^*)| - K(\tau - t^*)^\mu.$$

Since for any $\tau \in [t^*, t^* + \varepsilon]$ we have

$$|x(t^*)| - |x(\tau)| \leq K(\tau - t^*)^\mu,$$

it follows that

$$|x(\tau)| \geq |x(t^*)| - K(\tau - t^*)^\mu = \phi(\tau). \quad (7.52)$$

Noting that ε is defined by (7.51), we conclude from (7.52) that on the interval $[t^*, t^* + \varepsilon]$ the function $x(\tau)$ remains of the same sign. Therefore the relation

$$\begin{aligned} |y(t^* + \varepsilon) - y(t^*)| &= \left| \int_{t^*}^{t^* + \varepsilon} x(\tau) d\tau \right| = \int_{t^*}^{t^* + \varepsilon} |x(\tau)| d\tau \geq \int_{t^*}^{t^* + \varepsilon} \phi(\tau) d\tau \\ &= |x(t^*)| \varepsilon - \frac{K\varepsilon^{1+\mu}}{1+\mu} = K^{-1/\mu} \left(\frac{\mu}{1+\mu} \right) |x(t^*)|^{(1+\mu)/\mu} \end{aligned}$$

is valid. Since, however, the inequality

$$|y(t^* + \varepsilon) - y(t^*)| \leq 2||y||_C$$

is always fulfilled, it follows from the bound obtained that

$$|x(t^*)| \leq \left[2 \left(\frac{1+\mu}{\mu} \right) K^{1/\mu} ||y||_C \right]^{\mu/(1+\mu)}. \quad (7.53)$$

Now let both endpoints of the above-mentioned ε -neighborhood of the point t^* be located outside the interval $[a, b]$. Consider also the function

$$\phi_1(\tau) = \begin{cases} |x(t^*)| - |x(\tau)| \left(\frac{t^* - \tau}{t^* - a} \right)^\mu & \text{for } a \leq \tau \leq t^*, \\ |x(t^*)| - |x(\tau)| \left(\frac{\tau - t^*}{b - t^*} \right)^\mu & \text{for } t^* < \tau \leq b. \end{cases}$$

It is easy to verify that for any $\tau \in [a, b]$ the inequality

$$0 \leq \phi_1(\tau) \leq |x(\tau)|$$

is fulfilled. Therefore as above we have

$$|y(b) - y(a)| = \left| \int_a^b x(t) dt \right| = \int_a^b |x(t)| dt \geq \int_a^b \phi(\tau) d\tau = \frac{\mu}{1+\mu} (b-a) |x(t^*)|$$

Hence

$$|x(t^*)| \leq \frac{2}{b-a} \left(\frac{1+\mu}{\mu} \right) ||y||_C. \quad (7.54)$$

Thus if at least one of the endpoints of the ε -neighborhood is located within the interval $[a, b]$, the bound (7.53) is valid; otherwise, (7.54) is valid. While the inequalities were obtained for any t^* such that $|x(t^*)| \neq 0$, the bound

$$||x||_C \leq \max \left\{ \{ C_0^* ||y||_C; C_0^{**} ||y||_C^{\mu/(1+\mu)} \} \right\} \quad (7.55)$$

holds true, where

$$C_0^* = \frac{2}{b-a} \left(\frac{1+\mu}{\mu} \right), \quad C_0^{**} = \left[2 \left(\frac{1+\mu}{\mu} \right) K^{1/\mu} \right]^{\mu/(1+\mu)}.$$

For case $m=0$ the lemma is thus proved.

2. Now consider the bound

$$\|x^{(m-i)}\|_C \leq \max \left\{ C_i^* \|x^{(m-i-1)}\|_C; C_i^{**} \|x^{(m-i-1)}\|_C^{(1+\mu)/(1+i+\mu)} \right\}, \quad (7.56)$$

where $x^{(s)}$ is the s th derivative of function $x(t)$. This bound was obtained above for the case $i=0$. For the case $i=m$ it constitutes the assertion of the lemma (here we use the notation $x^{-1}(t)=y(t)$). We shall prove the validity (7.56) for $i=1, \dots, m$ by induction.

Let the bound (7.56) be valid for $i=k-1$. We show that it remains valid for $i=k$ as well. Indeed, since $x^{(m-k)}(t)$ is differentiable on $[a,b]$, we have

$$\sup_{t,\tau \in [a,b]} |x^{(m-k)}(t) - x^{(m-k)}(\tau)| \leq \|x^{(m-k+1)}\|_C |t - \tau|;$$

hence the function $x^{(m-k)}$ satisfies the Lipschitz condition of order $\mu=1$. Therefore utilizing (7.55) we obtain

$$\|x^{(m-k)}\|_C \leq \max \left\{ \frac{2^2}{b-a} \|x^{(m-k-1)}\|_C; 2 \|x^{(m-k+1)}\|_C^{1/2} \|x^{(m-k-1)}\|_C^{1/2} \right\}.$$

By the induction assumption we have

$$\|x^{(m-k+1)}\|_C \leq \max \left\{ C_{k-1}^* \|x^{(m-k)}\|_C; C_{k-1}^{**} \|x^{(m-k)}\|_C^{(k-1+\mu)/(k+\mu)} \right\}.$$

Combining these two inequalities, we have

$$\begin{aligned} \|x^{(m-k)}\|_C &\leq \max \left\{ \frac{2^2}{b-a} \|x^{(m-k-1)}\|_C; \right. \\ &2 \left[C_{k-1}^* \|x^{(m-k)}\|_C \right]^{1/2} \|x^{(m-k-1)}\|_C^{1/2}, \\ &2 \left[C_{k-1}^{**} \|x^{(m-k)}\|_C \right]^{(k+\mu-1)/2(\mu+k)} \|x^{(m-k-1)}\|_C^{1/2} \left. \right\} \end{aligned} \quad (7.57)$$

It follows from (7.57) that

$$\begin{aligned} \|x^{(m-k)}\|_C &\leq \max \left\{ \frac{2^2}{b-a} \|x^{(m-k-1)}\|_C; \frac{2}{b-a} C_{k-1}^* \|x^{(m-k-1)}\|_C; \right. \\ &\left. \left(4 C_{k-1}^{**} \|x^{(m-k-1)}\|_C \right)^{(\mu+k)/(\mu+k+1)} \right\}. \end{aligned}$$

Finally, taking the values C_k^* and C_k^{**} into account, we arrive at the inequality

$$\|x^{(m-k)}\|_C \leq \max \left\{ C_k^* \|x^{(m-k-1)}\|_C; C_k^{**} \|x_{(m-k-1)}\|_{(\mu+k)/(\mu+k+1)}^{(t)} \right\}.$$

For $k = m$ the inequality obtained is assertion of the lemma.

Proof of the Theorem. According to the iterated logarithm law the deviation between the empirical distribution function $F_\ell(x)$ and actual distribution function $F(x)$ satisfies with probability one the relation

$$\limsup_{\ell \rightarrow \infty} \left(\frac{2\ell}{\ln \ln \ell} \right)^{1/2} \sup_x |F_\ell(x) - F(x)| = 1.$$

Therefore for any ε there exists $\ell_0 = \ell(\varepsilon)$ such that simultaneously for all $\ell > \ell_0$ the inequality

$$\frac{\ell}{\ln \ln \ell} \|F_\ell(x) - F(x)\|_C^2 < 1 \quad (7.58)$$

is fulfilled with probability $1 - \varepsilon$.

Let $f_\ell(t)$ be the function that minimizes the functional (7.50), and let $f(t)$ be the desired density. Then

$$\begin{aligned} \frac{\ln \ln \ell}{\ell} W(f_\ell) &\leq R_\ell(f_\ell, F_\ell) \leq R_\ell(f, F_\ell) \\ &= \|F_\ell(x) - F(x)\|_C^2 + \frac{\ln \ln \ell}{\ell} W(f), \end{aligned}$$

from which we obtain

$$W(f_\ell) \leq W(f) + \|F_\ell(x) - F(x)\|_C^2 \frac{\ell}{\ln \ln \ell}. \quad (7.59)$$

Observe that starting with $\ell = \ell_0$, the inequality (7.58) is fulfilled with probability $1 - \eta$; hence starting with ℓ_0 , the inequality

$$W(f_\ell) \leq W(f) + 1 \quad (7.60)$$

is satisfied with probability $1 - \eta$. If the m th derivative of the desired density $f(t)$ satisfies the Lipschitz condition of order μ and the functional $W(f)$ is (7.49), then it follows from (7.60) that

$$\sup_{t,\tau} \frac{|f_\ell^{(m)}(t) - f_\ell^{(m)}(\tau)|}{|t - \tau|^\mu} \leq (W(f) + 1)^{1/2};$$

that is, the m th derivative of the function $f_\ell(t)$ satisfies with probability $1 - \eta$ the Lipschitz condition of order μ with the constant

$$K = (W(f) + 1)^{1/2}.$$

Therefore in view of the lemma, the inequality

$$\begin{aligned} \|f - f_\ell\|_C &\leq \max \left\{ C_m^* \left\| \int_a^b \theta(x-t)f_\ell(t) dt - F(x) \right\|_C; \right. \\ &\quad \left. C_m^{**} \left\| \int_a^b \theta(x-t)f_\ell(t) dt - F(x) \right\|_C^{\beta/(1+\beta)} \right\} \quad (\beta = m + \mu) \quad (7.61) \end{aligned}$$

is valid with probability $1 - \eta$. Multiplying both sides of inequality by

$$\left(\frac{\ell}{\ln \ln \ell} \right)^{\beta/2(1+\beta)},$$

we obtain

$$\begin{aligned} &\left(\frac{\ell}{\ln \ln \ell} \right)^{\beta/2(1+\beta)} \|f_\ell - f\|_C \\ &\leq \max \left\{ C_m^* \left(\sqrt{\frac{\ln \ln \ell}{\ell}} \right)^{1/(1+\beta)} \left[\sqrt{\frac{\ell}{\ln \ln \ell}} \left\| \int_a^b \theta(x-t)f_\ell(t) dt - F(x) \right\|_C \right]; \right. \\ &\quad \left. C_m^{**} \left[\sqrt{\frac{\ell}{\ln \ln \ell}} \left\| \int_a^b \theta(x-t)f_\ell(t) dt - F(x) \right\|_C \right]^{\beta/(1+\beta)} \right\}. \quad (7.62) \end{aligned}$$

Observe now that starting with ℓ_0 the inequality

$$\sqrt{\frac{\ell}{\ln \ln \ell}} \left\| \int_a^b \theta(x-t)f_\ell(t) dt - F(x) \right\|_C \leq 1 + \sqrt{W(f) + 1} \quad (7.63)$$

is fulfilled for all ℓ with probability $1 - \eta$. The inequality (7.63) follows from the triangle inequality

$$\begin{aligned} &\left\| \int_a^b \theta(x-t)f_\ell(t) dt - F(x) \right\|_C \\ &\leq \left\| \int_a^b \theta(x-t)f_\ell(t) dt - F_\ell(x) \right\|_C + \|F(x) - F_\ell(x)\|_C, \end{aligned}$$

the self-evident system of inequalities

$$\left\| \int_a^b \theta(x-t)f_\ell(t) dt - F_\ell(x) \right\|_C^2 \leq R_\ell(f_\ell, F_\ell) \leq R_\ell(f, F_\ell),$$

and the bound (7.58).

Taking (7.62) and (7.63) into account, we may assert that with probability $1 - \eta$ for all $\ell > \ell_0$ the inequality

$$\left(\frac{\ell}{\ln \ln \ell} \right)^{\beta/2(1+\beta)} \|f_\ell - f\|_C \leq \max \left\{ C_m^{***} \left(\frac{\ln \ln \ell}{\ell} \right)^{1/2(1+\beta)}; g \right\} \quad (7.64)$$

is fulfilled, where

$$\begin{aligned} C_m^{***} &= C_m^* (1 + \sqrt{1 + W(f)}), \\ g &= C_m^{**} \left(1 + \sqrt{1 + W(f)} \right)^{\beta/(1+\beta)}. \end{aligned}$$

Evidently, starting with some number ℓ_0 , the inequality

$$C_m^{***} \left(\frac{\ln \ln \ell}{\ell} \right)^{1/2(1+\beta)} < g$$

is satisfied. Thus starting with some ℓ_0 with probability $1 - \eta$ the inequality

$$\left(\frac{\ell}{\ln \ln \ell} \right)^{\beta/2(1+\beta)} \|f_\ell - f\|_C < g \quad (7.65)$$

will be fulfilled. Since for any $0 \leq \eta < 1$ there exists $\ell_0 = \ell(\eta)$ such that for all $\ell > \ell_0$ simultaneously the inequality (7.61) is fulfilled with probability $1 - \eta$, we have with probability 1

$$\limsup_{\ell \rightarrow \infty} \left(\frac{\ell}{\ln \ln \ell} \right)^{\beta/2(1+\beta)} \|f_\ell - f\|_C < g.$$

The theorem is proved.

7.9 CHOOSING A VALUE OF SMOOTHING (REGULARIZATION) PARAMETER FOR THE PROBLEM OF DENSITY ESTIMATION

The last section showed that if in the functional (7.50) one uses the value of regularization parameter γ_ℓ which is proportional to $\ln \ln \ell / \ell$, then one

obtains the almost optimal asymptotic rate of convergence in order of magnitude. However, for practical goals the asymptotic optimality is not enough. To get a good result, it is necessary to find a way to evaluate the best value of the regularization parameter for **a given fixed** amount of observations.

In this section we will consider some general way for finding a value of the regularization parameter γ_ℓ , which looks reasonable for a fixed amount of observations and provides the almost optimal rate of convergence when the number of observations tends to infinity.

In Chapter 1, Section 1.11, describing relations between the empirical distribution function and the actual distribution function we introduced Kolmogorov law, Smirnov law, and the law of iterated logarithm.

According to these laws, the specific measure $r(F_\ell, F)$ between the empirical distribution function $F_\ell(x)$ and the actual one $F(x)$ (different laws correspond to different measures) have distributions that are independent of both the actual distribution function $F(x)$ and the number ℓ (for sufficiently large ℓ).

Thus, according to the Kolmogorov law for sufficiently large ℓ the random variables

$$\xi = \sqrt{\ell} \sup_x |F(x) - F_\ell(x)|$$

have some fixed distribution function.

According to the Smirnov law for sufficiently large ℓ the random variables

$$\omega^2 = \ell \int (F(x) - F_\ell(x))^2 dF(x)$$

also have some fixed distribution function. Both of these distributions are unimodal functions.[†]

Let a_K be the median of the random variable ξ (one can find that for the one-dimensional case, $a_K \approx 0.6$) and let a_S be the median of the random variable ω^2 (for the one-dimensional case, $a_S \approx 0.05$).

Now let $p_{\gamma_\ell}(t)$ be an approximation to desired density which depends on the smoothing parameter γ_ℓ . The idea is to choose the value of the smoothing parameter γ_ℓ for which the corresponding distribution function

$$F_{\gamma_\ell}(x) = \int_{-\infty}^x f_{\gamma_\ell}(t) dt$$

satisfies a (chosen specific) statistical law in the best manner.

That means that if one uses Kolmogorov's law, then one has to choose a value γ_ℓ^* such that

$$\sup_x |F_{\gamma_\ell^*}(x) - F_\ell(x)| = \frac{a_K}{\sqrt{\ell}}. \quad (7.66)$$

[†]There are other laws for checking goodness of fit that are based on these two.

If one uses Smirnov's law, then one has to choose a value of the parameter γ_ℓ^* such that

$$\int (F_{\gamma_\ell}(x) - F_\ell(x))^2 f_{\gamma_\ell}(x) dx = \frac{a_s}{\ell}. \quad (7.67)$$

The value γ_ℓ^* satisfying (7.66) is the best fit of the regularization parameter for Kolmogorov's law.

The value γ_ℓ^* satisfying (7.67) is the best fit of the regularization parameter for Smirnov's law.

For the one-dimensional case, Eqs. (7.66) and (7.67) have simple expressions. Let

$$\bar{x}_1, \dots, \bar{x}_\ell$$

be an ordered array of observations x_1, \dots, x_ℓ . Then Eq. (7.66) can be rewritten in the form

$$\max_{1 \leq i \leq \ell} \max \left(\left| F_{\gamma_\ell^*}(\bar{x}_i) - \frac{i}{\ell} \right|, \left| F_{\gamma_\ell^*}(\bar{x}_i) - \frac{i-1}{4} \right| \right) = \sqrt{\ell} - \frac{1}{2}. \quad (7.68)$$

Equation (7.67) can be rewritten in the form

$$\sum_{i=1}^{\ell} \left(F_{\gamma_\ell^*}(\bar{x}_i) - \frac{i-0.5}{\ell} \right)^2 = a_s - \frac{1}{12\ell}. \quad (7.69)$$

One can show that using Eq. (7.69) for choosing the parameter of regularization, it is possible to achieve an almost optimal rate of convergence. Namely, let the density $p(t)$ satisfy the following conditions:

1. The unknown density is concentrated on $[0, \pi]$.
2. There exists the k th derivative ($k \geq 1$) of this density which has bounded variation on $[0, \pi]$.
3. The function $p(t)$ can be extended in the even manner to $[-\pi, \pi]$ and then periodically to the entire real axis so that the k th derivative of it will be continuous.

Any density of this type can be represented in the form of series with respect to orthogonal basis

$$\cos t, \dots, \cos mt, \dots$$

in $L_2(0, \pi)$:

$$p(t) = \frac{1}{\pi} + \frac{2}{\pi} \sum_{r=1}^{\infty} a_r^0 \cos rt,$$

where

$$a_r^0 = \int_0^\pi p(t) \cos rt dt, \quad r = 1, \dots$$

Let us consider the projection estimator $p_\ell(t)$ of the density obtained by the regularization methods — that is, the functions minimizing the functional

$$R_\ell(p, F_\ell) = \int_0^\pi \left(\int_0^x p(t) dt - F_\ell(x) \right)^2 dx + \gamma_\ell \int_0^\pi \left(p^{(k)}(t) \right)^2 dt$$

for a fixed $\gamma_\ell > 0$.

In the Section 7.5.2 we showed that $p_\ell(t)$ has the form

$$p(t) = \frac{1}{\pi} + \frac{2}{\pi} \sum_{r=1}^{\infty} \lambda_r a_r \cos rt,$$

where

$$\begin{aligned} c_r &= \frac{1}{\ell} \sum_{i=1}^{\ell} \cos rt_i, \quad r = 1, \dots, \\ \lambda_r &= \frac{1}{1 + \gamma_\ell r^{2(k+1)}}. \end{aligned}$$

The last equation defines regularization of the coefficients of expansion, which depend on the parameter γ_ℓ . The following theorem is true (Vapnik, Markovich, and Stefanyuk, 1991).

Theorem 7.5. *If the regularization parameter is chosen to satisfy Smirnov goodness-of-fit equation, then for $k \geq 1$ there exists a constant c such that the inequality*

$$P \left\{ \limsup_{\ell \rightarrow \infty} \ell^{(k+0.5)/(2k+3)} \|p_\ell(t) - p(t)\|_{L_2} < c \right\} = 1$$

holds true.

The asymptotic bounds obtained in this theorem are slightly worse than the best possible $\ell^{k/(2k+1)}$: The largest difference is achieved for $k = 1$ where $\ell^{0.3}$ instead of $\ell^{\frac{1}{3}}$. However, this rule for choosing the regularization constant γ_ℓ has no free parameters and it performs well in practice (Markovich, 1989).

7.10 ESTIMATION OF THE RATIO OF TWO DENSITIES

In this section we consider the problem of estimation of the ratio of two densities

$$f(x) = \frac{p_1(x)}{p_2(x)}$$

using two sets of data x_1, \dots, x_n and $\bar{x}_1, \dots, \bar{x}_\ell$ drawn randomly and independently in accordance with $p_1(x)$ and $p_2(x)$. We consider the case where f belongs to the set of functions with bounded variation.

In Example 2 of Section 7.1 we have considered the problem of estimating the ratio of two densities as a problem of solving the equation

$$\int_0^x f(t) dF^{(2)}(t) = F^{(1)}(t), \quad (7.70)$$

where $F^{(1)}(x)$ and $F^{(2)}(x)$ are distribution functions corresponding to the densities $p_1(x)$ and $p_2(x)$. We have to solve this equation in the situation where $F^{(1)}(x)$ and $F^{(2)}(x)$ are unknown but we are given the data

$$\begin{aligned} &x_1, \dots, x_n, \\ &\bar{x}_1, \dots, \bar{x}_\ell. \end{aligned} \quad (7.71)$$

Here we require that

$$\lim_{\ell \rightarrow \infty} \frac{\ell}{n} = c, \quad 0 < c < \infty.$$

Let us use these data to estimate the unknown distribution functions

$$F_n^{(1)}(x) = \frac{1}{n} \sum \theta(x - x_i),$$

$$F_\ell^{(2)}(x) = \frac{1}{\ell} \sum \theta(\bar{x} - \bar{x}_i).$$

These approximations of the unknown distribution functions determine an approximation to the right-hand side of operator equation (7.70) and an approximation to the operator

$$A_\ell f = \int_0^x f(x) dF_\ell^{(2)}(x). \quad (7.72)$$

To estimate the densities ratio we will minimize the functional

$$R_{\gamma_\ell}^*(f, F_n, A_\ell) = \rho_{E_2}^2(A_\ell f, F_n) + \gamma_\ell W(f), \quad (7.73)$$

where F_n and A_ℓ are the approximations described above.

For simplicity consider the case where densities have a bounded support $[a, b]$ and where the ratio of two densities is a continuous function with bounded variation that possesses $m \geq 0$ derivatives, satisfying the Lipschitz condition order γ

$$\sup_{x,y \in [a,b]} \frac{|f^{(m)}(x) - f^{(m)}(y)|}{|x - y|^\gamma} < \infty.$$

(If $m \geq 1$, then the function on $[a,b]$ possesses bounded variation. The bounded variation requirement is essential for case $m = 0$.) Consider the $C(a,h)$ norm

$$\|f\| = \sup_{x \in [a,b]} |f(x)|.$$

Let D be a set of functions with bounded variation $V_a^b(f)$, possessing m derivatives, satisfying the Lipschitz condition of order γ .

Consider also the functional

$$W(f) = \left(V_a^b(f) + \max_{0 \leq m \leq k} \sup_{a \leq x \leq b} |f^{(m)}(x)| + \sup_{x,y \in [a,b]} \frac{|f^{(m)}(x) - f^{(m)}(y)|}{|x-y|^\gamma} \right)^2$$

Let $f_{\ell,n}$ be a sequence of minima of functional (7.73). Below we show that the conditions (7.34) and (7.35) on the regularization parameter γ_ℓ which guarantee convergence (in probability and almost surely) of density will guarantee convergence in probability and convergence almost surely of the estimate of the ratio of the densities as well.

Indeed consider the difference

$$\begin{aligned} (A_\ell f)(x) - (Af)(x) &= \int_0^x f(t) d(F_\ell(t) - F(t)) \\ &= f(x)(F_\ell(x) - F(x)) - \int_0^x (F_\ell(x) - F(x)) df(x). \end{aligned}$$

Therefore,

$$\|A_\ell f - Af\|_{E_2} \leq \|f\|_{E_1} \|F_\ell - F\|_{E_2} + \|F_\ell - F\|_{E_2} V_a^b(f) \leq \|F_\ell - F\|_{E_2} W^{1/2}(f).$$

From this inequality and the definition of the norm of the operator we have

$$\|A_\ell - A\| = \sup_f \frac{\|A_\ell f - Af\|_{E_2}}{W^{1/2}(f)} \leq \|F_\ell - F\|_{E_2}. \quad (7.74)$$

According to Theorem 7.3 the solution f_γ of the operator equation obtained on the basis of the regularization method possesses the following properties: For any $\epsilon > 0$, $C_1 > 0$, and $C_2 > 0$, there exists γ_0 such that for any $\gamma_\ell < \gamma_0$ the inequality

$$P\{\rho_{E_1}(f_{\ell,n}, f) > \epsilon\} \leq P\{\rho_{E_2}(F_n, F) > C_1\sqrt{\gamma_\ell}\} + P\{\|A_\ell - A\| > C_2\sqrt{\gamma_\ell}\}$$

holds true. Therefore for our special case taking into account the Kolmogorov inequality we have

$$\begin{aligned} P\{\rho_{E_1}(f_{\ell,n}, f) > \epsilon\} &\leq P\{\sup_{0 \leq x \leq 1} |F_n(x) - F(x)| > C_1\sqrt{\gamma_\ell}\} + P\{\sup_{0 \leq x \leq 1} |F_\ell(x) - F(x)| > \frac{C_2}{2}\sqrt{\gamma_\ell}\} \\ &\leq 2 \left(\exp\{-2\gamma_\ell\ell C_1^2\} + 2 \exp\{-\gamma_\ell\ell C_2^2\} \right). \end{aligned}$$

From this inequality we find (as in the density estimation case) that conditions (7.34) and (7.35) imply convergence in probability and convergence almost surely for convergence of our solution to the desired one.

7.10.1 Estimation of Conditional Densities

Now we derive an estimator $P(y|x)$ of conditional densities and solve the equation

$$\int_0^x \int_0^y p(y/x) dF(x) dy = F(x, y)$$

using the data

$$(x_1, y_1), \dots, (x_\ell, y_\ell)$$

by minimizing the functional

$$\begin{aligned} R(p) = & \int \left[\int_0^y \int_0^x p(y'|x') d \left(\frac{1}{\ell} \sum_{i=1}^{\ell} \theta(x' - x_i) \right) dy' \right. \\ & \left. - \frac{1}{\ell} \sum_{i=1}^{\ell} \theta(x - x_i) \theta(y - y_i) \right]^2 dx dy + \gamma_\ell \int |\text{grad } p(y|x)|^2 dx dy \end{aligned} \quad (7.75)$$

Consider the case where $x \in (0, \tau)$ and $y \in (0, \tau)$. We are looking for the minimum of functional (7.75) in the form

$$p(y|x) = \frac{1}{\pi} + \sum_k^\infty a_{0,k} \cos ky + \sum_{m,k}^\infty a_{m,k} \cos mx \cos ky. \quad (7.76)$$

We shall search for an approximation to the minimum of the functional in the form of a finite sum:

$$p(y|x) = \frac{1}{\pi} + \sum_k^M a_{0,k} \cos ky + \sum_{m=1}^N \sum_{k=1}^M a_{m,k} \cos mx \cos ky. \quad (7.77)$$

The minimum of (7.75) in the set (7.77) is obtained for a function whose expansion coefficients $a_{m,k}$, $m = 0, 1, \dots, N$, $k = 1, \dots, M$, satisfy the following system of algebraic equations:

$$\sum_{m=0}^N B(m, r) a_{m,k} + \gamma_\ell h_r k^2 (r^2 + k^2) a_{r,k} = D(r, k), \quad (7.78)$$

$$h_r = \begin{cases} \pi^2/2 & \text{for } r = 0, \\ (\pi/2)^2 & \text{for } r \neq 0, \end{cases} \quad (7.79)$$

$$r = 0, \dots, N, \quad k = 1, \dots, M,$$

where

$$\begin{aligned} B(m, r) &= \frac{1}{\ell^2} \sum_{i,j=1}^{\ell} \left(\frac{\pi - \max(x_i x_j)}{\pi} \right) \cos mx_i \cos rx_j, \\ D(r, k) &= \frac{1}{\ell^2} \sum_{i,j=1}^{\ell} \left(\frac{\pi - \max(x_i x_j)}{\pi} \right) \cos rx_i \cos ky. \end{aligned} \quad (7.80)$$

In accordance with (7.78) and (7.79) to estimate the $(N+1)M$ coefficients $a_{m,k}$ of the series (7.77), it is sufficient to solve M times the system of linear algebraic equations (7.78).

Using the estimator of the conditional probability density (7.77), it is easy to calculate the regression

$$\begin{aligned} r_{\ell}(x) &= \int y p(y|x) dy \\ &= \frac{\pi}{2} - 2 \sum_{p=0}^{\lfloor \frac{M-1}{2} \rfloor} \frac{a_{0,(2p+1)}}{(2p+1)^2} - 2 \sum_{r=1}^N \sum_{p=0}^{\lfloor \frac{M-1}{2} \rfloor} \frac{a_{r,(2p+1)}}{(2p+1)^2} \cos rx. \end{aligned} \quad (7.81)$$

7.11 ESTIMATION OF THE RATIO OF TWO DENSITIES ON THE LINE

Now let \mathbf{x} be a random vector of dimension n . We shall estimate the conditional probability density on the line

$$\mathbf{x} - \mathbf{x}_0 = et \quad (7.82)$$

passing through point x_0 , where e is the unit vector defining the line.

For this purpose, along with line (7.82) we shall consider its orthogonal complement with respect to space X . We define every point in X by both the number t (the location of its projection on the line (7.82) relative to the point x_0) and the $(n-1)$ -dimensional vector u . Thus any point in X space is defined by the pair $x_i = (t_i, u_i)$, and any point on the line is defined by the pair $\mathbf{x} = (t, u_0)$. We introduce two conditional distribution functions $F(t|u_0)$ and $F(y, t|u_0)$. The equation determining the conditional density $p(y|t, u_0)$ on the line is u_0

$$\int_a^y \int_0^t p(\bar{y}|\bar{t}, u_0) dF(\bar{t}|u_0) d\bar{y} = F(y, t|u_0). \quad (7.83)$$

We shall solve this equation when the conditional distribution functions are unknown but a sample of pairs

$$(y_1, x_1), \dots, (y_{\ell}, x_{\ell})$$

is given. We rewrite these pairs in the equivalent form

$$(y_1, t_1, u_1), \dots, (y_\ell, t_\ell, u_\ell).$$

Consider the following approximations to the conditional distribution functions

$$\begin{aligned} F_\ell(t|u) &= \sum_{i=1}^{\ell} \tau_i(u) \theta(t - t_i), \\ F_\ell(y, t|u) &= \sum_{i=1}^{\ell} \tau_i(u) \theta(t - t_i) \theta(y - y_i), \end{aligned} \quad (7.84)$$

where coefficients $\tau_i(u)$ are defined using the kernel estimator

$$\tau_i(u) = \frac{g_{\sigma_\ell}(|u - u_i|)}{\sum_{i=1}^{\ell} g_{\sigma_\ell}(|u - u_i|)}$$

(σ_ℓ depends on number of observations ℓ).

It is known (Stute, 1986) that if the kernel $g_{\sigma_\ell}(|u - u_i|)$, $u \in R^{n-1}$, is such that

$$g_{\sigma_\ell}(|u - u_i|) = g\left(\frac{|u - u_i|}{\sigma_\ell}\right),$$

where

$$g(a) = g(a^1, \dots, a^{n-1}) = 1, \quad \text{if } -\frac{1}{2} \leq a^i \leq \frac{1}{2} \text{ for } i = 1, \dots, n-1$$

and zero elsewhere and σ_ℓ satisfies the property that $\sigma_\ell \rightarrow 0$ in such a way that

$$\sum_{\ell \geq 1} \exp\{-\rho \ell \sigma_\ell^{n-1}\} < \infty \quad \forall \rho > 0,$$

then almost for all u we have

$$D_1(u) = \sup_t |F(t|u) - F_\ell(t|u)| \xrightarrow{\ell \rightarrow \infty} 0,$$

$$D_2(u) = \sup_{y,t} |F(y, t|u) - F_\ell(y, t|u)| \xrightarrow{\ell \rightarrow \infty} 0$$

with probability one.

The problem is thus to estimate the solution of Eq. (7.83) based on the approximations (7.84). We assume that the conditional density has a bounded

support $0 \leq t \leq \pi$, $0 \leq y \leq \pi$. To obtain the solution we shall minimize the functional

$$\begin{aligned} R(p) = & \int_0^\pi \int_0^\pi \left[\int_0^y \int_0^t p(\hat{y}|t, u_0) d\left(\sum_{i=1}^\ell \tau_i(u_0) \theta(\hat{t} - t_i) \right) d\hat{y} d\hat{t} \right. \\ & \left. - \sum_{i=1}^\ell \tau_i(u_0) \theta(t - t_i) \theta(y - y_i) \right]^2 dt dy + \gamma_\ell \int |\text{grad } p(y|t, u_0)|^2 dt dy \end{aligned} \quad (7.85)$$

in the set of functions

$$p(y|t, u_0) = \frac{1}{\pi} + \sum_{k=1}^M a_{0,k} \cos ky + \sum_{m=1}^N \sum_{k=1}^M a_{m,k} \cos mt \cos ky. \quad (7.86)$$

The minimum of (7.85) in the set (7.86) is obtained for a function whose expansion coefficients $a_{m,k}$, $m = 0, 1, \dots, N$, $k = 1, \dots, M$, satisfy the following system of algebraic equations:

$$\sum_{m=0}^N B(m, r) a_{m,k} + \gamma_\ell h_r k^2 (r^2 + k^2) a_{r,k} = D(r, k), \quad (7.87)$$

$$\begin{aligned} h_r = & \begin{cases} \pi^2/2 & \text{for } r = 0, \\ (\pi/2)^2 & \text{for } r \neq 0, \end{cases} \\ r = & 0, \dots, N, \quad k = 1, \dots, M, \end{aligned} \quad (7.88)$$

where

$$B(m, r) = \sum_{i,j=1}^\ell \left(\frac{\pi - \max(t_i, t_j)}{\pi} \right) \tau_i(u_0) \tau_j(u_0) \cos mt_i \cos rt_j, \quad (7.89)$$

$$D(r, k) = \sum_{i,j=1}^\ell \left(\frac{\pi - \max(t_i, t_j)}{\pi} \right) \tau_i(u_0) \tau_j(u_0) \cos rt_i \cos ky_i.$$

In accordance with (7.87) to estimate the $(N+1)M$ coefficients $a_{m,k}$ of the series (7.86) it is sufficient to solve M times the system of linear algebraic equations (7.87).

Using the estimator of the conditional probability density (7.86) it is easy to calculate the regression estimator

$$\begin{aligned} r_\ell(t|u_0) = & \int y p(y|t, u) dy \\ = & \frac{\pi}{2} - 2 \sum_{p=0}^{\lfloor \frac{M-1}{2} \rfloor} \frac{a_{0,(2p+1)}}{(2p+1)^2} - 2 \sum_{r=1}^N \sum_{p=0}^{\lfloor \frac{M-1}{2} \rfloor} \frac{a_{r,(2p+1)}}{(2p+1)^2} \cos rt. \end{aligned} \quad (7.90)$$

7.12 ESTIMATION OF A CONDITIONAL PROBABILITY ON A LINE

To estimate a conditional probability $P_{z_0}(w = 1|t)$ (the probability of the first class given position on the line (7.82)) we use the following equation

$$\int_0^t P(w = 1|\hat{t}, u_0) dF(\hat{t}|u_0) = F(w = 1, t|u_0), \quad (7.91)$$

where $F(\hat{t}|u_0) = P\{t \leq \hat{t}|u_0\}$ and $F(w = 1, \hat{t}|u_0) = P\{w = 1, t \leq \hat{t}|u_0\}$.

We shall solve this equation when the functions $F(t|u_0)$ and $F(w = 1, t|u_0)$ are unknown and a sample of pairs

$$(w_1, x_1), \dots, (w_\ell, x_\ell)$$

is given. Rewrite these data in the form

$$(w_1, t_1, u_1), \dots, (w_\ell, t_\ell, u_\ell).$$

Consider the estimators

$$\begin{aligned} F_\ell(t|u_0) &= \sum_{i=1}^\ell \tau_i(u_0) \theta(t - t_i) \\ F_\ell(w = 1, t|u_0) &= \sum_{i=1}^\ell \delta(w_i) \tau_i(u_0) \theta(t - t_i), \end{aligned} \quad (7.92)$$

where $\delta(w_i)$ is the indicator of class: $\delta(w_i) = 1$ if vector x_i belongs to the first class and equal to zero otherwise. Let $t \in [0, \pi]$. We determine the solution (7.91) by minimizing the functional

$$\begin{aligned} R\{P(w = 1|t, u_0)\} \\ = \int_0^\pi \left(\int_0^t P(w = 1|t, u_0) d\left(\sum_{i=1}^\ell \tau_i(u_0) \theta(t - t_i) \right) - \sum_{i=1}^\ell \delta(w_i) \tau_i(u_0) \theta(t - t_i) \right)^2 dt \\ + \gamma \int_0^\pi \left(P^{(k)}(w = 1|t, u_0) \right)^2 dt \end{aligned} \quad (7.93)$$

on the set

$$P(w = 1|t, u_0) = a_0 + \sum_{i=1}^N a_i \cos it.$$

The minimum of the functional (7.93) is attained when the coefficients a_i satisfy the system of algebraic equations

$$\sum_{m=0}^N B(m, r) a_m + \gamma r^2 a_r = D(r), \quad r = 0, \dots, N, \quad (7.94)$$

where

$$\begin{aligned} B(m, r) &= \sum_{i,j=1}^{\ell} \left(\frac{\pi - \max(t_i t_j)}{\pi} \right) \tau_i(u_0) \tau_j(u_0) \cos mt_i \cos rt_j, \\ D(r) &= \sum_{i,j=1}^{\ell} \left(\frac{\pi - \max(t_i t_j)}{\pi} \right) \tau_i(u_0) \tau_j(u_0) \cos rt_i \delta(w_i). \end{aligned} \quad (7.95)$$

The last two sections described methods for estimating conditional densities (conditional probabilities) on the line. However, they did not discuss the problem of how to choose the line in order to achieve the best performance. In Chapter 11, we will come back to this problem and show that in a sufficiently general situation one can define the trajectory passing through the point of interest along which the conditional density (conditional probability) changes rapidly. We will estimate our functions along this trajectory.

8

ESTIMATING THE VALUES OF FUNCTION AT GIVEN POINTS

This chapter considers a new setting of the learning problem; the problem of estimating the values of a function at given points of interest. This setting of the problem leads to a new type of inference, the so-called *transductive inference* which is different from inductive inference. In contrast to the inductive inference where one uses given empirical data to find the approximation of a functional dependency (the inductive step) and then uses the obtained approximation to evaluate the values of a function at the points of interest (the deductive step), we will try to estimate the values of a function at the points of interest in one step.

8.1 THE SCHEME OF MINIMIZING OVERALL RISK

In the case of small sample size[†]

$$(y_1, x_1), \dots, (y_\ell, x_\ell) \tag{8.1}$$

we distinguish between two estimation problems:

1. Estimation of the functional dependence $y = \phi(x)$ in the class $f(x, a)$, $a \in A$.
2. Estimation of values of the function $y = \phi(x)$ at the given points

$$x_{\ell+1}, \dots, x_{\ell+k} \tag{8.2}$$

[†]For the problem of estimating the function on the basis of the set of functions $f(x, a)$ $a \in A$, the sample size ℓ is considered to be "small" if the ratio ℓ/h is small, say $\ell/h < 20$, where h is the VC dimension of the set.

using a function from the class $f(x, a), a \in A$. The data (8.2) are generated by the same generator that generates vectors x , from the training data (8.1).

It may seem that the problem of estimating the values of a function at given points (8.2) is not a very profound one. There exists a "natural" way to solve it: Based on the available empirical data (8.1), one can find the approximation $y = f(x, a^*)$ to the desired function $y = \phi(x)$, and then one can use this approximation to evaluate the values of the unknown function at the points (8.2):

$$y_i = f(x_i, a^*), \quad i = \ell + 1, \dots, \ell + k;$$

that is, one can obtain a solution of the second problem by using a solution of the first one.

However, this way for estimating values of a function is often not the best, since here a solution of a relatively simple problem (namely, estimating k numbers (the values of the function)) becomes dependent on the solution of a substantially more complex problem (namely, estimating a function (which is estimating the values in the continuum of points containing these k points)).

The problem is how to utilize the information about the data (8.2) for estimating its values using the set of functions $f(x, a), a \in A$.

It should be noted that in practice usually it is necessary to determine the values of the function at given points rather than to determine the functional dependence itself. As a rule (which is always valid for the problem of pattern recognition), the functional dependence is utilized only to determine the value of a function at certain desired points.

Thus we distinguish between two kinds of estimation problem: *estimation of a function* and *estimation of the values of a function at given points*.

In Chapter 1 we formalized the statement of the problem of estimation of functional dependence by means of a scheme of minimizing the expected risk. In this section we shall formalize the statement of the problem of estimating the functional values at given points using a scheme that will be called the scheme of *minimizing the overall risk functional*.

It is assumed that a set

$$x_1, \dots, x_\ell, x_{\ell+1}, \dots, x_{\ell+k}, \tag{8.3}$$

containing $\ell + k$ vectors (a *complete sample of vectors*) is given. These vectors are i.i.d. according to some distribution function. There exists a function $y = \phi(x)$ that assigns a number y to each vector x in the set (8.3). Thus for $\ell + k$ vectors (8.3), $\ell + k$ values

$$y_1, \dots, y_\ell, y_{\ell+1}, \dots, y_{\ell+k} \tag{8.4}$$

are defined; ℓ vectors x_i are randomly selected from the set (8.3) for which the corresponding realizations of y_i are indicated. The set of pairs

$$(x_1, y_1), \dots, (x_\ell, y_\ell) \quad (8.5)$$

thus formed is called the *training sample*. The remaining set of vectors

$$x_{\ell+1}, \dots, x_{\ell+k} \quad (8.6)$$

is called the *working sample*.

Below we consider two settings of the problem of estimating the values of the function $\phi(x)$ at the points of interest (8.6) using the training data (8.5).

Setting 1. Based on the elements of the training and the working samples and on the given set of functions $f(x, a)$, $a \in A$ ($\phi(x)$ does not necessarily belong to this set), it is required to find a function $f(x, a^*)$ that minimizes with a preassigned probability $1 - \eta$ the overall risk of forecasting the values of the function $y_i = \phi(x_i)$ on the elements of the working sample—that is, which yields with probability $1 - \eta$ a value of the functional

$$R_\Sigma(\alpha) = \frac{1}{k} \sum_{i=1}^k \rho(y_{\ell+i}, f(x_{\ell+i}, \alpha)) \quad (8.7)$$

close to the minimal one. In (8.7), $\rho(y, f(x, a))$ is some measure of discrepancy between y and $f(x, a)$, say

$$\rho(y_{\ell+i}, f(x_{\ell+i}, \alpha)) = (y_{\ell+i} - f(x_{\ell+i}, \alpha))^2.$$

Consider another formulation of this problem, to be referred to as *Setting 2*.

Setting 2. Let the probability distribution function $P(x, y)$ be given on the set of pairs (X, Y) (it can be an infinite set). We select from this set, randomly and independently, ℓ pairs

$$(y_1, x_1), \dots, (y_\ell, x_\ell), \quad (8.8)$$

which form the training sequence. Next, in the same manner we choose k additional pairs

$$(y_{\ell+1}, x_{\ell+1}), \dots, (y_{\ell+k}, x_{\ell+k}). \quad (8.9)$$

It is required to obtain an algorithm A which, based on the training sequence (8.5) and the working sequence (8.6), will choose a function

$$f(x, \alpha_A) = f(x, \alpha_A(x_1, y_1; \dots; x_\ell, y_\ell; x_{\ell+1}, \dots, x_{\ell+k}))$$

that yields the value of the functional

$$R(A) = \int \frac{1}{k} \sum_{i=\ell+1}^{\ell+k} \rho(y_i, f(x_i, \alpha_A)) dP(x_1, y_1) \cdots dP(x_{\ell+k}, y_{\ell+k})$$

close to the minimal one.

The following theorem that connects these two settings is valid.

Theorem 8.1. If for some algorithm A it is proved that for Setting 1 with probability $1 - \eta$ the deviation between the risk on the training data and the working sample does not depend on the composition of the complete sample and does not exceed ε , then with the same probability for Setting 2 the deviation between the analogous values of the risks does not exceed ε .

Proof. Denote

$$C_A(x_1, y_1; \dots; x_{\ell+k}, y_{\ell+k}) = \left| \frac{1}{\ell} \sum_{i=1}^{\ell} \rho(y_i, f(x_i, \alpha_A)) - \frac{1}{k} \sum_{i=\ell+1}^{\ell+k} \rho(y_i, f(x_i, \alpha_A)) \right|.$$

Consider Setting 2 of the problem, and compute the probability of deviation from zero by an amount greater than ε of the quantity $C_A(x_1, y_1; \dots; x_{\ell+k}, y_{\ell+k})$:

$$P = \int_{XY} \theta [C_A(x_1, y_1; \dots; x_{\ell+k}, y_{\ell+k}) - \varepsilon] dP(x_1, y_1) \cdots dP(x_{\ell+k}, y_{\ell+k}).$$

Let T_p , $p = 1, \dots, (\ell+k)!$ be the permutation operator for the sample $(x_1, y_1); \dots; (x_{\ell+k}, y_{\ell+k})$. Then the equality

$$\begin{aligned} P &= \int_{XY} \theta [C_A(x_1, y_1; \dots; x_{\ell+k}, y_{\ell+k}) - \varepsilon] dP(x_1, y_1) \cdots dP(x_{\ell+k}, y_{\ell+k}) \\ &= \int_{XY} \left\{ \frac{1}{(\ell+k)!} \sum_{p=1}^{(\ell+k)!} \theta [C_A(T_p(x_1, y_1; \dots; x_{\ell+k}, y_{\ell+k})) - \varepsilon] \right\} \\ &\quad dP(x_1, y_1) \cdots dP(x_{\ell+k}, y_{\ell+k}) \end{aligned}$$

is valid. The expression in braces is the quantity estimated in Setting 1. It does not exceed $1 - \eta$. We thus obtain

$$P \leq \int_{XY} (1 - \eta) dP(x_1, y_1) \cdots dP(x_{\ell+k}, y_{\ell+k}) = 1 - \eta.$$

The theorem is proved.

Below we shall consider the problem of estimating the values of a function at given points in Setting 1. However, by means of Theorem 8.1 all the results obtained are valid for the case of Setting 2.

The terminology used in this chapter pertains to estimating values of a function. However, all the results obtained were valid in the more general case when a realization of the sample (8.4) is determined by the conditional probability (rather than by the function $y = \phi(x)$); and it is required on the basis of random realizations at points (8.5) to forecast, by means of the functions $f(x, a)$, $a \in A$, realizations at some other points (8.6).

8.2 THE METHOD OF STRUCTURAL MINIMIZATION OF OVERALL RISK

We solve the problem of estimating the values of a function at given points by using the method of structural risk minimization. In the following two sections we obtain bounds on the rate of uniform relative deviation of the mean values in two subsamples. Using these bounds, we construct bounds on the overall risk, uniform over the class $f(x, a)$, $a \in A$, based on the values of the empirical risks. These bounds are analogous to those which were utilized in Chapter 6 when constructing a structural minimization of the expected risk.

We shall demonstrate that for a set of indicator functions of VC dimension h (for the problem of pattern recognition) the additive bound

$$R_{\Sigma}(\alpha) \leq R_{\text{emp}}(\alpha) + \Omega(\ell, k, h, \eta) \quad (8.10)$$

is valid with probability $1 - \eta$, while for the set of arbitrary functions of VC-dimension h with probability $1 - \eta$ the multiplicative bound

$$R_{\Sigma}(\alpha) \leq R_{\text{emp}}(\alpha)\Omega^*(\ell, k, h, \eta) \quad (8.11)$$

is valid.

Now if one defines the structure

$$S_1 \subset \dots \subset S_q$$

on the set of functions $f(x, a)$, $a \in A$, then it is possible by minimizing the right-hand side of the equality (8.10) (or (8.11)) to find an element S_* and a function $f(x, a^*)$ for which the guaranteed minimum for the bound of the overall risk is attained. Using the functions $f(x, \alpha_{\text{emp}}^*)$, the values $y_i = f(x_i, a_i^*)$ are computed at the points of the working sample. Outwardly this scheme does not differ at all from the one considered in Chapter 6.

However, in the scheme of structural minimization of the overall risk, a special feature determines the difference between solutions of problems of

estimating a function and those of estimating values of a function at given points. This has to do with the need to construct the structure a priori. This requirement has different meanings in the case of estimating functions and of estimating values of functions.

For the problem of estimating functions, it means that knowing the class of functions $f(x, a), a \in A$, and the domain of definition of a function, it is necessary to define a structure on $f(x, a), a \in A$. For the problem of estimating functional values, it amounts to determining a structure on $f(x, a), a \in A$, knowing the set of functions and the complete sample

$$x_1, \dots, x_\ell, x_{\ell+1}, \dots, x_{\ell+k}. \quad (8.12)$$

The difference stems from the fact that for a complete sample (8.12) the set of functions $f(x, a), a \in A$, is decomposed into sets of *equivalence classes*. This set can be investigated, and the structure on $f(x, a), a \in A$, can be defined on equivalence classes, producing a more meaningful ordering principle than the one in the case of estimating functions.

For example, the set of indicator functions on the complete sample (8.12) is decomposed into a finite number of equivalence classes. Two indicator functions are equivalent on a complete sample if they subdivide this sample into subsamples in the same manner (i.e., take the same values on (8.12)). In this case, one can define a structure on a finite number of equivalence classes rather than on the initial (possibly infinite) set of functions.

8.3 BOUNDS ON THE UNIFORM RELATIVE DEVIATION OF FREQUENCIES IN TWO SUBSAMPLES

This section finds for a bound on the uniform relative deviation of frequencies in two subsamples. For the problem of minimizing the overall risk in the class of indicator functions, this bound plays the same role as the bound on uniform relative deviation of frequencies from their probabilities played in the problem of minimizing the expected risk. To state the theorem we shall introduce for a given set of indicator functions $f(x, a), a \in A$, and any given set

$$x_1, \dots, x_{\ell+k}$$

the finite number of equivalence classes

$$N_{\ell+k} = N(x_1, \dots, x_{\ell+k}).$$

Observe that the number of equivalence classes on the complete sample is bounded using the growth function as follows:

$$N_{\ell+k} \leq \exp \left\{ G^\Lambda(\ell + k) \right\}.$$

Let our set of $\ell + k$ vectors x consist of elements of two types: m elements of type a and $\ell + k - m$ elements of type b . We select randomly ℓ elements from this set. The probability that among the selected elements there are r elements of type a equals

$$P(r, \ell + k, \ell, m) = \frac{C_m^r C_{\ell+k-m}^{\ell-r}}{C_{\ell+k}^{\ell}} \quad (8.13)$$

Thus with probability (8.13) the frequency of elements of type a in the selected group is r/ℓ , and hence the corresponding frequency in the remaining group is $(m-r)/k$.

The probability that the frequency of elements a in the first group deviates from the frequency of elements a in the second group by the amount exceeding ε is equal to

$$P \left\{ \left| \frac{r}{\ell} - \frac{m-r}{k} \right| > \varepsilon \right\} = \sum_r \frac{C_m^r C_{\ell+k-m}^{\ell-r}}{C_{\ell+k}^{\ell}} = \Gamma_{\ell,k}(\varepsilon, m),$$

where the summation is taken over the values of r such that

$$\left| \frac{r}{\ell} - \frac{m-r}{k} \right| > \varepsilon, \quad \max(0, m-k) \leq r \leq \min(\ell, m).$$

We define the function

$$\Gamma_{\ell,k}(\varepsilon) = \max_m \Gamma_{\ell,k} \left(\sqrt{\frac{m}{\ell+k}} \varepsilon, m \right).$$

This function can be tabulated with a computer.

Denote now by $\nu_0(\alpha)$ the frequency of classification error on the set $x_1, \dots, x_{\ell+k}$ when using the decision rule $f(x, a)$. Denote also by $\nu(\alpha)$ the frequency of errors on the set x_1, \dots, x_ℓ and by $\nu_\Sigma(\alpha)$ the frequency of errors on the set $x_{\ell+1}, \dots, x_{\ell+k}$. Clearly,

$$\nu_0(\alpha) = \frac{k}{\ell+k} \nu_\Sigma(\alpha) + \frac{\ell}{\ell+k} \nu(\alpha).$$

The following theorem on uniform relative deviation of frequencies in the two subsamples is valid.

Theorem 8.2. Let the set of decision rules $f(x, a), a \in A$ on the complete set of vectors have $N_{\ell+k}$ equivalence classes. Then the probability that the relative size of deviation for at least one rule in $f(x, a), a \in A$ exceeds ε is bounded by

$$P \left\{ \sup_{\alpha \in \Lambda} \frac{|\nu(\alpha) - \nu_\Sigma(\alpha)|}{\sqrt{\nu_0(\alpha)}} > \varepsilon \right\} < N_{\ell+k} \Gamma_{\ell,k}(\varepsilon). \quad (8.14)$$

Here we use the convention

$$\frac{|\nu(\alpha) - \nu_\Sigma(\alpha)|}{\sqrt{\nu_0(\alpha)}} = 0 \quad \text{for } \nu(\alpha) = \nu_\Sigma(\alpha) = \nu_0(\alpha) = 0.$$

Proof. Since number of equivalence classes is finite, the inequality

$$P \left\{ \sup_{\alpha \in \Lambda} \frac{|\nu(\alpha) - \nu_\Sigma(\alpha)|}{\sqrt{\nu_0(\alpha)}} > \varepsilon \right\} < N \sup_{\alpha \in \Lambda} P \left\{ \frac{|\nu(\alpha) - \nu_\Sigma(\alpha)|}{\sqrt{\nu_0(\alpha_0)}} > \varepsilon \right\}$$

is valid. The second term on the right-hand side of this inequality is bounded using the function $\Gamma_{\ell,k}(\varepsilon)$ (in Chapter 4, Section 4.13, we obtained the bound of this function for $k = P$). Indeed,

$$\begin{aligned} & P \left\{ \frac{|\nu(\alpha) - \nu_\Sigma(\alpha)|}{\sqrt{\nu_0(\alpha)}} > \varepsilon \right\} \\ &= P \{ |\nu(\alpha) - \nu_\Sigma(\alpha)| > \varepsilon \sqrt{\nu_0(\alpha)} \} \\ &= P \left\{ |\nu(\alpha) - \nu_\Sigma(\alpha)| > \varepsilon \sqrt{\frac{m}{\ell+k}} \right\} = \Gamma_{\ell,k} \left(\varepsilon \sqrt{\frac{m}{\ell+k}}, m \right), \end{aligned}$$

and by definition,

$$\Gamma_{\ell,k} \left(\varepsilon \sqrt{\frac{m}{\ell+k}}, m \right) \leq \Gamma_{\ell,k}(\varepsilon)$$

The theorem is proved.

Below, a uniform in $f(x, a), a \in A$, bound on the frequency of errors in the working sample will be required. We shall derive it using Theorem 8.2. We bound the right-hand side of (8.14) by the quantity η . We thus arrive at the inequality

$$\ln N_{\ell+k} + \ln \Gamma_{\ell,k}(\varepsilon) \leq \ln \eta,$$

the smallest solution of which (with respect to ε) we denote by \mathcal{E} .

Using this solution we can rewrite (8.14) in the equivalent form: With probability $1 - \eta$ the inequality

$$\nu_\Sigma(\alpha) \leq \nu(\alpha) + \frac{\mathcal{E}^2 k}{2(\ell+k)} + \mathcal{E} \sqrt{\nu(\alpha) + \left(\frac{k\mathcal{E}}{2(\ell+k)} \right)^2} \quad (8.15)$$

is valid simultaneously for all a . We shall utilize this inequality when constructing algorithms for structural minimization of the risk in the class of indicator functions.

8.4 A BOUND ON THE UNIFORM RELATIVE DEVIATION OF MEANS IN TWO SUBSAMPLES

When deriving a bound on the uniform relative deviation of the means in two subsamples we shall assume that on the complete sample

$$x_1, \dots, x_\ell, x_{\ell+1}, \dots, x_{\ell+k}$$

the condition

$$\sup_{\alpha \in \Lambda} \frac{\sqrt[p]{\frac{1}{\ell+k} \sum_{i=1}^{\ell+k} (y_i - f(x_i, \alpha))^2 p}}{\frac{1}{\ell+k} \sum_{i=1}^{\ell+k} (y_i - f(x_i, \alpha))^2} \leq \tau, \quad p > 2 \quad (8.16)$$

is fulfilled for a set of real functions $f(x, \alpha), \alpha \in A$, where y_i is a value of the realization of (8.4).

The condition (8.16) conveys some prior information concerning possible large deviations on the complete sample $x_1, \dots, x_{\ell+k}$. This condition is analogous to the condition considered in Chapter 5, Section 5.8.

In the same manner as in Chapter 5, we introduce the function

$$R_1(\alpha) = \int \sqrt{\nu\{(y - f(x, \alpha))^2 > t\}} dt, \quad p > 2,$$

where $\nu\{\rho(y, f(x, \alpha)) > t\}$ is the ratio of the number of points in the complete sample $x_1, \dots, x_{\ell+k}$ for which the condition $\rho(y, f(x, \alpha)) > t$ is fulfilled on realizations of (8.4) to the total number of points $\ell + k$. For the function $R_1(\alpha)$ similarly to the function $R(\alpha)$ (cf. Chapter 5) the relation

$$R_1(\alpha) \leq a(p) \sqrt[p]{\frac{1}{\ell+k} \sum_{i=1}^{\ell+k} (y_i - f(x_i, \alpha))^2 p} \quad (8.17)$$

holds true, where

$$a(p) = \sqrt[p]{\frac{1}{2} \left(\frac{p-1}{p-2} \right)^{p-1}}.$$

Denote

$$R(\alpha) = \frac{1}{\ell+k} \sum_{i=1}^{\ell+k} \rho(y_i, f(x_i, \alpha)) = \frac{\ell}{\ell+k} R_{\text{emp}}(\alpha) + \frac{k}{\ell+k} R_{\Sigma}(\alpha). \quad (8.18)$$

The following theorem is valid.

Theorem 8.3. Let the condition (8.16) be satisfied and let the set of indicators $\theta(f(x, \alpha) + \beta)$ of real-valued functions $f(x, \alpha), \alpha \in A$, have $N_{\ell+k}$ equivalence classes on a complete set of x . Then the bound

$$P \left\{ \sup_{\alpha \in \Lambda} \frac{|R_{\Sigma}(\alpha) - R_{\text{emp}}(\alpha)|}{R(\alpha)} > \tau a(p) \varepsilon \right\} < N_{\ell+k} (\ell + k) \Gamma_{\ell,k}(\varepsilon) \quad (8.19)$$

is valid.

Proof. Let the number of different separations of the set of pairs

$$(y_1, x_1), \dots, (y_{\ell+k}, x_{\ell+k})$$

using indicator functions

$$I(\alpha, \beta) = \theta \{(y - f(x, \alpha))^2 - \beta\}$$

not exceed

$$N^* = N_{\ell+k} (\ell + k)$$

and therefore according to Theorem 8.2 the bound

$$P \left\{ \sup_{\alpha \in \Lambda} \frac{|\nu(\alpha, \beta) - \nu_{\Sigma}(\alpha, \beta)|}{\sqrt{\nu_0(\alpha, \beta)}} > \varepsilon \right\} < N^* \Gamma_{\ell,k}(\varepsilon) \quad (8.20)$$

is valid. (Here $\nu(\alpha, \beta)$ is the frequency of the event $\{\rho(y, f(x, \alpha)) > \beta\}$ computed for the training sequence, $\nu_{\Sigma}(A_{\alpha, \beta})$ is the frequency of the event $\{\rho(y, f(x, \alpha)) > \beta\}$ computed for the working sample $x_{\ell+1}, \dots, x_{\ell+k}$, and $\nu_0(A_{\alpha, \beta})$ is the frequency of event $(A_{\alpha, \beta})$ computed for the complete sample $x_1, \dots, x_{\ell+k}$ via the realization of (8.4).)

We show that the validity of (8.20) implies the validity of the inequality

$$P \left\{ \sup_{\alpha} \frac{|R_{\Sigma}(\alpha) - R_{\text{emp}}(\alpha)|}{R_1(\alpha)} > \varepsilon \right\} < N^* \Gamma_{\ell,k}(\varepsilon). \quad (8.21)$$

For this purpose we write the expression

$$I = \sup_{\alpha} \frac{|R_{\Sigma}(\alpha) - R_{\text{emp}}(\alpha)|}{R_1(\alpha)}$$

in the form of a Lebesgue integral

$$I = \sup_{\alpha} \lim_{n \rightarrow \infty} \frac{\frac{1}{n} \sum_{i=1}^{\infty} \left| \nu_{\Sigma} \left\{ (y - f(x, \alpha))^2 > \frac{i}{n} \right\} - \nu_{\text{emp}} \left\{ (y - f(x, \alpha))^2 > \frac{i}{n} \right\} \right|}{R_1(\alpha)}$$

Now let the inequality

$$\frac{\left| \nu_{\Sigma} \left\{ (y - f(x, \alpha))^2 > \frac{i}{n} \right\} - \nu_{\text{emp}} \left\{ (y - f(x, \alpha))^2 > \frac{i}{n} \right\} \right|}{\sqrt{\nu \left((y - f(x, \alpha))^2 > \frac{i}{n} \right)}} \leq \varepsilon$$

be valid. In that case

$$\frac{|R_{\Sigma}(\alpha) - R_{\text{emp}}(\alpha)|}{R_1(\alpha)} \leq \lim_{n \rightarrow \infty} \frac{\varepsilon \sqrt{\nu \left((y - f(x, \alpha))^2 > \frac{i}{n} \right)}}{R_1(\alpha)} = \varepsilon.$$

The validity of (8.20) implies that (8.21) holds.

To complete the proof it remains only to utilize the inequalities (8.16) and (8.17). Indeed

$$\begin{aligned} P \left\{ \sup_{\alpha} \frac{|R_{\Sigma}(\alpha) - R_{\text{emp}}(\alpha)|}{R(\alpha)} > \tau a(p) \varepsilon \right\} &\leq P \left\{ \sup_{\alpha \in \Lambda} \frac{|\nu(\alpha, \beta) - \nu_{\Sigma}(\alpha, \beta)|}{\sqrt{\nu_0(\alpha, \beta)}} > \varepsilon \right\} \\ &< N^* \Gamma_{\ell, k}(\varepsilon). \end{aligned}$$

The theorem is proved.

We shall now obtain a uniform bound for the risk on the working sample. For this purpose we bound the right-hand side of (8.19) by the quantity η . We thus arrive at the inequality

$$\ln N^* + \ln \Gamma_{\ell, k}(\varepsilon) \leq \ln \eta. \quad (8.22)$$

Denote by \mathcal{E} the smallest solution with regard to ε for this inequality.

Taking the representation (8.18) into account, we obtain from (8.19) that the inequality

$$R_{\Sigma}(\alpha) < \frac{\left(1 + \tau a(p) \frac{\ell}{\ell + k} \mathcal{E} \right)}{\left(1 - \tau a(p) \frac{\ell}{\ell + k} \mathcal{E} \right)_+} R_{\text{emp}}(\alpha), \quad (8.23)$$

where

$$(u)_+ = \max(u, 0),$$

is valid with probability $1 - \eta$.

This inequality will be utilized in the course of constructing algorithms for a structural minimization of the overall risk. Below we shall confine our discussion to a class of functions linear in parameters

$$f(x, \alpha) = \sum_{r=1}^{n-1} \alpha_r \phi_r(x) + \alpha_0.$$

The capacity of this class of functions is equal to n .

8.5 ESTIMATION OF VALUES OF AN INDICATOR FUNCTION IN A CLASS OF LINEAR DECISION RULES

Let the complete sample

$$x_1, \dots, x_\ell, x_{\ell+1}, \dots, x_{\ell+k} \quad (8.24)$$

be given. On this sample the set of decision rules is decomposed into a finite number N of equivalence classes F_1, \dots, F_N . Two decision rules $F(x, a^*)$ and $F(x, \alpha^{**})$ fall into the same equivalence class if they subdivide the sample (8.24) into two subsamples in the same manner. Altogether, $N^A(x_1, \dots, x_{\ell+k})$ subdivisions of the sample (8.24) into two classes by means of the rules $f(x, a)$, $a \in A$, are possible, and thus there exist $N^A(x_1, \dots, x_{\ell+k})$ equivalence classes.

Recall that by the definition of the entropy and the growth function (cf. Chapter 4, Section 4.9) the inequality

$$\ln N^A(x_1, \dots, x_{\ell+k}) \leq G^A(\ell + k)$$

is valid. For linear decision rules in a space of dimension n the following bound is valid (Chapter 4, Section 4.9):

$$N^* \leq \exp \left\{ n \left(\ln \frac{\ell + k}{n} + 1 \right) \right\}.$$

Thus on the complete sample (8.24) the set of linear decision rules $f(x, a)$, $a \in A$, is decomposed on N^* equivalence classes F_1, \dots, F_{N^*} .

Observe that the equivalence classes are not of equal size. Some of them contain more decision rules than others. We assign to each equivalence class a quantity that characterizes the fraction of linear decision rules they encompass. Such a quantity can be constructed. Indeed, assign to each function

$$f(x, \alpha) = \theta \left(\sum_{r=1}^n \alpha_r \phi_r(x) \right)$$

a directional vector (Fig. 8.1)

$$\alpha = (\alpha^1, \dots, \alpha^n)^T, \quad |\alpha| = 1.$$

Then in the space of parameters α a unit sphere corresponds to the set of all hyperplanes; and to each equivalence class F_i there corresponds a distinct region on the surface of the sphere. (The set of N equivalence classes subdivides the sphere into N regions.) The ratio of the area corresponding to the region \mathcal{L}_i to the area of the sphere \mathcal{L} characterizes the fraction of functions belonging to an equivalence class relative to all possible linear decision rules.

Now order the equivalence classes in decreasing order of $\pi_i = \mathcal{L}_i/\mathcal{L}$ and introduce the following structure:

$$S_1 \subset S_2 \subset \dots \subset S_q, \quad (8.25)$$

where the element S_r contains only those equivalence classes that satisfy

$$\frac{\mathcal{L}_i}{\mathcal{L}} \geq c_r, \quad c_1 > c_2 > \dots > c_q = 0.$$

We have thus constructed a structure in which each element S_p possesses an extremal property: For a given number of equivalence classes it contains the maximal share of all decision rules. However, it is difficult to compute the value $\mathcal{L}_i/\mathcal{L}$ and thus to form the structure (8.25). Therefore we shall consider another characteristic of the size of equivalence classes which is similar to $\mathcal{L}_i/\mathcal{L}$ in its meaning and can be obtained in practice.

Denote by ρ_r the value of the distance between the convex hulls of the two classes into which are placed vectors of the complete sample allocated to different classes by the decision rules belonging to F_r , and assign to the equivalence class F_r the number

$$\pi(F_r) = \frac{\rho_r}{D}, \quad (8.26)$$

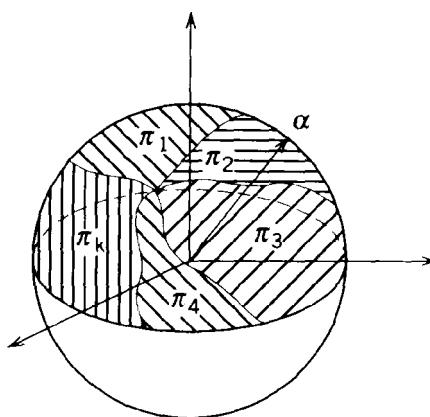


FIGURE 8.1. To each equivalence class F_i there corresponds a distinct region on the surface of the sphere.

where $D/2$ is the radius of the smallest sphere containing the set (8.24); that is,

$$\frac{D}{2} = \min_{x^*} \max_{x_1, \dots, x_{\ell+k}} ||x_i - x^*||.$$

Now we define a structure

$$S_1 \subset S_2 \subset \dots \subset S_n \quad (8.27)$$

on the equivalence classes; here S_r contains only those equivalence classes F_i such that

$$\begin{aligned} \pi(F_i) &> \frac{1}{\sqrt{r-1}} & \text{for } r < n, \\ \pi(F_i) &\geq 0 & \text{for } r > n, \quad r \geq 2. \end{aligned} \quad (8.28)$$

The set S_1 in (8.27) is empty.

To construct a method of structural risk minimization for the overall risk on the structure (8.27) we shall bound the number N_r of equivalence classes belonging to the element of the structure S_r .

The following theorem is valid.

Theorem 8.4. *The number of equivalence classes in S_r is bounded by*

$$N_r < \exp \left\{ r \left(\ln \frac{\ell+k}{r} + 1 \right) \right\}, \quad (8.29)$$

where

$$r = \min \left(n, \left[\frac{D^2}{\rho^2} \right] + 1 \right), \quad (8.30)$$

n is the dimensionality of the space, and [a] is the integer part of number a.

Proof. Observe that the number N_r equals the maximal number of subdivisions of the sample

$$x_1, \dots, x_{\ell+k}$$

into two classes such that the distance between their convex hulls exceeds $D/\sqrt{r-1}$; that is,

$$\rho > \frac{D}{\sqrt{r-1}} = \rho_r \quad (8.31)$$

According to Theorem 4.3 (Chapter 4, Section 4.9) the number of such decision rules does not exceed

$$G^\Lambda(\ell+k) < \exp \left\{ r \left(\ln \frac{\ell+k}{r} + 1 \right) \right\},$$

where r is the maximal number of points in the sample for which an arbitrary subdivision into two classes satisfies (8.31). Observe that if the condition (8.31) is fulfilled, then the subdivision is evidently carried out by means of a hyperplane; therefore obviously

$$r \leq n,$$

where n is the dimension of the space.

Now consider r points

$$\mathbf{x}_1, \dots, \mathbf{x}_r$$

and 2^r possible subdivisions of these points into two subsets

$$T_1, \dots, T_{2^r}.$$

Denote by $\rho_r(T_i)$ the distance between the convex hulls of vectors belonging to distinct subsets under subdivision T_i .

The fact that (8.31) is fulfilled for any T , can be written as

$$\min_i \rho(T_i) > \rho_r.$$

Then the number r does not exceed the maximal number of vectors such that the inequality

$$H(r) = \max_{\mathbf{x}_1, \dots, \mathbf{x}_r} \min_i \rho(T_i) \geq \frac{D}{\sqrt{r-1}} \quad (8.32)$$

is still fulfilled. It follows from symmetry considerations that the maximal r is attained where the vectors $\mathbf{x}_1, \dots, \mathbf{x}_r$ are located at the vertices of a regular $(r-1)$ -dimensional simplex inscribed in a sphere of radius $D/2$, and T_i is a subdivision into (a) two subsimplices of dimension $(r/2-1)$ for even r and (b) two subsimplices of dimensions $(r-1)/2$ and $(r-3)/2$ for odd r . Therefore elementary calculations show that

$$H(r) = \begin{cases} \frac{D}{\sqrt{r-1}} & \text{for even } r, \\ \frac{D}{\sqrt{r-1}} \sqrt{\frac{r^2}{r^2-1}} & \text{for odd } r. \end{cases}$$

For $r > 10$ the quantities

$$\frac{1}{\sqrt{r-1}} \quad \text{and} \quad \frac{1}{\sqrt{r-1}} \sqrt{\frac{r^2}{r^2-1}}$$

are close to each other (they differ by the amount less than 0.01). Thus we take

$$H(r) = \frac{D}{\sqrt{r-1}}. \quad (8.33)$$

(A bound from the above on $H(r)$ would have been the expression

$$H(r) < \frac{D}{\sqrt{r-1.01}}, \quad r > 10. \quad \left. \right)$$

It follows from the inequalities (8.32) and (8.33) that for integer r we obtain

$$r < \left[\frac{D^2}{\rho^2} \right] + 1.$$

Finally, taking into account that the subdivision is done by means of a hyperplane (i.e., $r \leq n$), we obtain

$$r \leq \min \left(\left[\frac{D^2}{\rho^2} \right] + 1, n \right). \quad (8.34)$$

Consequently in view of Theorem 4.3 we have

$$N_r \leq \exp \left\{ r \left(\frac{\ell+k}{r} + 1 \right) \right\}.$$

The theorem is thus proved.

It follows from Theorem 8.2 and Theorem 8.4 that with probability $1 - \eta$ simultaneously for all decision rules in S_r the inequality

$$\nu_{\Sigma}(\alpha) \leq \nu(\alpha) + \frac{k\mathcal{E}^2}{2(\ell+k)} + \mathcal{E} \sqrt{\nu(\alpha) + \left(\frac{k\mathcal{E}}{2(\ell+k)} \right)^2} = R(\alpha, r) \quad (8.35)$$

is satisfied, where \mathcal{E} is the smallest solution of the inequality

$$r \left(\ln \frac{\ell+k}{r} + 1 \right) + \ln \Gamma_{\ell,k}(\mathcal{E}) \leq \ln \eta.$$

The method of structural minimization of the overall risk consists of indexing the working sample by means of the rule $f(x, a, \cdot)$ which minimizes the functional (8.35) with respect to r and $a \in A$. Let the minimum be equal to $R(\alpha_{\text{emp}}^*, r_*)$. For such an indexing procedure the assertion

$$P \{ \nu_{\Sigma}(\alpha_{\text{emp}}^*) < R(\alpha_{\text{emp}}^*, r_*) \} > 1 - n\eta$$

is valid.

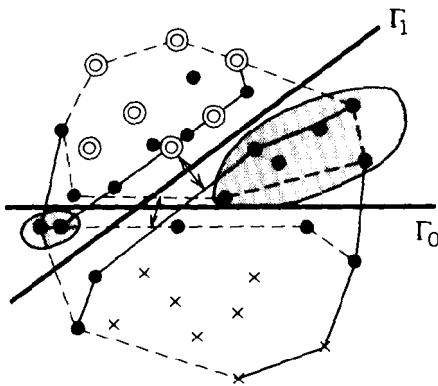


FIGURE 8.2. Two approaches to the problem of estimating values of the function at the given points. Using inductive inference we construct the separating hyperplane Γ_0 , and using transductive inference we construct the separating hyperplane Γ_1 .

Consider an example that illustrates the difference between solving the problem of classifying vectors in the working sample using the method of minimizing the overall risk and using a decision rule that minimizes the empirical risk for a training sequence.

In Fig. 8.2, vectors of the first class of the training sequence are denoted by crosses, and vectors of the second class are denoted by small circles. Dots represent vectors of the working sample.

A solution of this problem within the framework of minimizing the expected risk consists in constructing a subdividing hyperplane that will ensure the minimal probability of error. Let the solution be chosen among hyperplanes that subdivide the vectors of the training sequence without error. In this case the minimal guaranteed probability of error is ensured by the optimal subdividing hyperplane (the one that is the farthest from the elements of the training sequence). The vectors that are located on different sides of the hyperplane Γ_0 are assigned to different classes. This determines the solution of the problem using the method of minimizing the empirical risk. A solution of the problem using the method of minimizing the overall risk is determined by the hyperplane Γ_1 , which maximizes the distance between the convex hulls of the subdivided sets. Vectors located on one side of the hyperplane belong to the first class, and those on the other side of the hyperplane belong to the second class.

The points of the working sample that are classified by the hyperplanes Γ_0 and Γ_1 in a different manner are shaded in Fig. 8.1.

8.6 SAMPLE SELECTION FOR ESTIMATING THE VALUES OF AN INDICATOR FUNCTION

We have seen that the solution of the problem of estimating the values of an indicator function at given points using the method of structural minimization

of the overall risk leads to results that are different from those obtained from the classification of vectors of the working sample

$$x_{\ell+1}, \dots, x_{\ell+k} \quad (8.36)$$

by means of a decision rule $f(x, \alpha_{\text{emp}})$ that minimizes the empirical risk on the elements of the training sequence

$$(\omega_1, x_1), \dots, (\omega_\ell, x_\ell). \quad (8.37)$$

This result was obtained because the complete sample

$$x_1, \dots, x_\ell, x_{\ell+1}, \dots, x_{\ell+k} \quad (8.38)$$

consisted of a small number of elements whose special location could be studied; it is related to a specific method of ordering the class of decision rules $f(x, a), a \in A$.

The method of ordering actually determined the difference in classification. Thus the geometry of vectors in the complete sample (8.38) predetermined the possibility of a more accurate solution of the problem of estimating the values of a function at given points.

If this is indeed the case, then the question arises: Is it possible, by excluding a few elements from the complete sample (8.38) (i.e., by changing the geometry of the vectors of the complete sample in space), to affect the construction of the structure on the class of decision rules in order to increase the guaranteed number of correct classifications of the elements in the working sample? It turns out that this is possible.[†]

We now present the idea of selection of a complete *sample*. Consider, along with the set X of vectors in the complete sample (8.38),

$$H'_{\ell+k} = \sum_{p=0} C_{\ell+k}^p$$

distinct subsets

$$X_1, \dots, X_{H'_{\ell+k}}$$

obtained from (8.38) by excluding at most t vectors. Now let the training sequence (8.37) and the working sample (8.36) be defined on the initial set of vectors (8.38). The training and working samples induce on each one of the sets $X_1, \dots, X_{H'_{\ell+k}}$ its own training and working subsamples.

Consider $H'_{\ell+k}$ problems of estimating values of a function at given points. Each one of these problems is determined by a training sequence

$$\overbrace{x_1, \omega_1; \dots; \overbrace{x_i, \omega_i; \dots; \overbrace{x_j, \omega_j; \dots; x_\ell, \omega_\ell}}$$

[†] We note that in the case of estimating an indicator function the selection of the training sample does not lead to a decrease in the estimation of the minimal guaranteed risk.

and a working sample

$$x_{\ell+1}, \dots, \widehat{x_{\ell+\ell}}, \dots, x_{\ell+k}$$

($\widehat{x, \omega}$ denotes that the element x, ω is excluded from the sequence).

For each problem, in accordance with its complete sample

$$x_1, \dots, \widehat{x_i}, \dots, \widehat{x_j}, \dots, \widehat{x_{\ell+\ell}}, \dots, x_{\ell+k}$$

we shall determine the equivalence classes of linear decision rules. We define a structure on the equivalence classes, utilizing the principle of ordering according to relative distance considered in the preceding section.

It follows from Theorem 8.2 and Theorem 8.4 that with probability $1 - \eta$ in each problem (separately) the inequality

$$\begin{aligned} \nu_{\Sigma}(\alpha_{\text{emp}}^r) &\leq \nu(\alpha_{\text{emp}}^r) + \frac{(k - k_{\text{exc}})\mathcal{E}^2}{2(\ell + k - k_{\text{exc}} - \ell_{\text{exc}})} \\ &+ \mathcal{E} \sqrt{\nu(\alpha_{\text{emp}}^r) + \left(\frac{(k - k_{\text{exc}})\mathcal{E}}{2(\ell + k - k_{\text{exc}} - \ell_{\text{exc}})} \right)^2} \end{aligned} \quad (8.39)$$

is valid for the rule $f(x, \alpha_{\text{emp}}^r)$ minimizing the empirical risk in S_r , where \mathcal{E} is the smallest solution of the equation

$$r \left(\ln \frac{\ell + k - k_{\text{exc}} - \ell_{\text{exc}}}{r} + 1 \right) + \ln \Gamma_{\ell - \ell_{\text{exc}}, k - k_{\text{exc}}}(\mathcal{E}) \leq \ln \eta. \quad (8.40)$$

In (8.39) and (8.40) the following notation is used: ℓ_{exc} is the number of elements excluded from the training sequence, and k_{exc} is the number of elements excluded from the working sample.

Simultaneously for the r th elements of the structure for all $H_{\ell+k}^{k_{\text{exc}}+\ell_{\text{exc}}}$ problems the inequality

$$\begin{aligned} \nu_{\Sigma}^{(i)}(\alpha_{\text{emp}}^r) &\leq \nu^{(i)}(\alpha_{\text{emp}}^r) + \frac{(k - k_{\text{exc}}^{(i)})(\mathcal{E}^{(i)})^2}{2(\ell + k - k_{\text{exc}}^{(i)} - \ell_{\text{exc}}^{(i)})} \\ &+ \mathcal{E}^{(i)} \sqrt{\nu^{(i)}(\alpha_{\text{emp}}^r) + \left(\frac{(k - k_{\text{exc}}^{(i)})\mathcal{E}^{(i)}}{2(\ell + k - k_{\text{exc}}^{(i)} - \ell_{\text{exc}}^{(i)})} \right)^2} \end{aligned} \quad (8.41)$$

is fulfilled with probability $1 - \eta$, where $\mathcal{E}^{(i)}$ are the smallest solutions of the equations

$$r \left(\ln \frac{\ell + k - k_{\text{exc}}^{(i)} - \ell_{\text{exc}}^{(i)}}{r} + 1 \right) + \ln H_{\ell+k}^i + \ln \Gamma_{\ell - \ell_{\text{exc}}^{(i)}, k - k_{\text{exc}}^{(i)}}(\mathcal{E}^i) \leq \ln \eta \quad (8.42)$$

where the total number of excluded elements varies from 1 to t . In (8.41) and (8.42) the following notation is used: $k_{\text{exc}}^{(i)}$ and $\ell_{\text{exc}}^{(i)}$ are the numbers of elements in the training and the working samples omitted from (8.37) and (8.36) when forming the i th problem, and $\nu_{\Sigma}^{(i)}(\alpha_{\text{emp}}^r)$ and $\nu^{(i)}(\alpha_{\text{emp}}^r)$ are the frequencies of erroneous classification of the working and the training samples in the i th problem. Multiply each of the inequalities (8.41) by $k - k_{\text{exc}}^{(i)}$. This will yield for each problem a bound on the number of errors m_i in $k - k_{\text{exc}}^{(i)}$ elements of the working sample:

$$\begin{aligned} m_i \leq & \left[\nu^{(i)}(\alpha_{\text{emp}}^r) + \frac{(k - k_{\text{exc}}^{(i)})(\mathcal{E}^{(i)})^2}{2(\ell + k - k_{\text{exc}}^{(i)} - \ell_{\text{exc}}^{(i)})} \right. \\ & \left. + \mathcal{E}^{(i)} \sqrt{\nu^{(i)}(\alpha_{\text{emp}}^r) + \left(\frac{(k - k_{\text{exc}}^{(i)})\mathcal{E}^{(i)}}{2(\ell + k - k_{\text{exc}}^{(i)} - \ell_{\text{exc}}^{(i)})} \right)^2} \right] (k - k_{\text{exc}}^{(i)}). \end{aligned} \quad (8.43)$$

If the number of vectors excluded from the working sequences were the same for all problems and equal to k_{exc} , then the best guaranteed solution of the problem of classifying $k - k_{\text{exc}}$ vectors in the working sample would be determined by the inequality (the problem) for which the bound on the number of errors in the $k - k_{\text{exc}}$ elements of the working sample is the smallest. However, the number of vectors excluded from the working sample is not the same for different problems. Therefore we shall consider as the best solution the one that maximizes the number of correct classifications of the elements of the working sample—that is, the one that minimizes the quantity[†]

$$\begin{aligned} R(r, i) = & \left[\nu(\alpha_{\text{emp}}^r) + \frac{(k - k_{\text{exc}}^{(i)})(\mathcal{E}^{(i)})^2}{2(\ell + k - k_{\text{exc}}^{(i)} - \ell_{\text{exc}}^{(i)})} \right. \\ & \left. + \mathcal{E}^{(i)} \sqrt{\nu^{(i)}(\alpha_{\text{emp}}^r) + \left(\frac{(k - k_{\text{exc}}^{(i)})\mathcal{E}^{(i)}}{2(\ell + k - k_{\text{exc}}^{(i)} - \ell_{\text{exc}}^{(i)})} \right)^2} \right] (k - k_{\text{exc}}^{(i)}) + k_{\text{exc}}^{(i)} \end{aligned} \quad (8.44)$$

(the number of errors plus the number of vectors excluded from the working sample).

Now by enumeration over r and k_{exc} we shall determine vectors that should be excluded to guarantee the largest number of correctly classified vectors in the working sample. The problem of minimizing the functional (8.44) with

[†]Here one can introduce different costs for error and refusal to classify elements in the working set.

respect to r and k_{exc} is quite difficult computationally. Its exact solution requires a large number of enumerations. However, by using certain heuristic methods, one can achieve a satisfactory solution in a reasonable amount of time.

Observe that in the course of selection of a complete sample, the elements are picked for both the training sample and for those of the working sample.

A selection of elements of the working sample allows us to increase the total number of correctly classified vectors at the expense of declining to classify certain elements.

Up until now we have assumed that the space on which the structure is constructed is fixed. However, the procedure of ordering with respect to relative distances may be carried out in any subspace E_m of the initial space E_n . Moreover, the minimal values of the corresponding bounds need not be obtained on the initial space E_n . This fact yields the possibility of achieving a more refined minimum for the bound on the risk by means of additional minimization over subspace.

8.7 ESTIMATION OF VALUES OF A REAL FUNCTION IN THE CLASS OF FUNCTIONS LINEAR IN THEIR PARAMETERS

Now we extend the methods of estimating values of indicator functions considered in the preceding sections to the estimating of values of a real function in a class of functions linear in their parameters.[†] For this purpose we shall determine equivalence classes of linear (in parameters) functions on a complete sample, define a structure on these classes, and implement the method of structural risk minimization.

Let a complete sample

$$x_1, \dots, x_\ell, x_{\ell+1}, \dots, x_{\ell+k} \quad (8.45)$$

and a set of linear (in parameters) functions $f(x, a)$, $a \in A$, be given. We shall assign to each function $f(x, a^*)$ in this set a one-parameter family (in the parameter β) of decision rules

$$f_{a^*}(\beta) = \theta(f(x, a^*) + \beta), \quad \beta \in (-\infty, \infty). \quad (8.46)$$

As the parameter β varies from $-\infty$ to ∞ , the family (8.46) forms a sequence of dichotomies (subdivisions into two classes) of the set of vectors (8.45): It starts with the dichotomy for which the first class is empty and the second class consists of the entire set of vectors

$$[\emptyset; \{x_1, \dots, x_{\ell+k}\}]$$

[†] Note that this does not mean that an unknown relationship is described by the function linear in its parameters. This means that we will approximate an unknown relationship by these functions.

(for $\beta = -\infty$), and it ends with the dichotomy for which the first class contains the entire set (8.45) and the second class is empty:

$$[\{x_1, \dots, x_{\ell+k}\}; \emptyset]$$

(for $\beta = \infty$). Thus for each function $f(x, a)$ a sequence of dichotomies

$$[\emptyset; \{x_1, \dots, x_{\ell+k}\}]; \dots; [\{x_1, \dots, x_{\ell+k}\}; \emptyset] \quad (8.47)$$

can be constructed. In accordance with this sequence of dichotomies we shall subdivide the set of functions $f(x, a), a \in A$ into a finite number of equivalence classes. Two functions $f(x, \alpha_1)$ and $f(x, \alpha_2)$ fall into the same equivalence class F_i if they form the same sequence of dichotomies (8.47).

Now assign to each equivalence class a number $\pi(F_i)$ that is equal to the fraction of all functions belonging to it, and then arrange the equivalence classes in the order of decreasing values of $\pi(F_i)$:

$$F_1, \dots, F_N, \quad \pi(F_1) \geq \pi(F_2) \geq \dots \geq \pi(F_N). \quad (8.48)$$

Utilizing this ordering (8.48), one can construct a structure on the set of the equivalence classes

$$S_1 \subset S_2 \subset \dots \subset S_n.$$

The element S_r contains those equivalence classes F_i for which

$$\pi(F_i) > c_r.$$

One can define the fraction of functions belonging to an equivalence class in the case of sets of linear functions in the same manner as the fraction of linear decision rules was determined. Now assign to each linear function a vector of direction cosines. Then the surface of the input sphere in the space of dimension n will correspond to the set of all functions, and a particular region on this sphere will correspond to each equivalence class (Fig. 8.1). The ratio of the area of a single-out region to the area of the surface of the sphere will determine the fraction of functions belonging to an equivalence class among the entire set of functions.

In practice, however, it is difficult to compute the characteristic $\pi(F_i)$ directly. Therefore, in the same manner as in Section 8.5, we shall consider another characteristic of the size of an equivalence class. For each function

$$f(x, \alpha) = \sum_{i=1}^n \alpha_i \phi_i(z)$$

we define a directional vector $\alpha / \|\alpha\|$. Each equivalence class, F_m is characterized by the number

$$\rho_m = \min_{z_i, z_j} \sup_a \left| (z_i - z_j)^T \left| \frac{\alpha}{\|\alpha\|} \right| \right|, \quad i \neq j,$$

where the minimum is taken over all the vectors of the complete sample, and the supremum is taken over all directional vectors of a given equivalence class.

We now form the following structure:

$$S_1 \subset \dots \subset S_n,$$

The functions for which

$$\begin{aligned}\pi^2(F) &= \left[\frac{\rho^2}{D^2} \right] > \frac{1}{d-1} && \text{for } d < n, \\ \pi^2(F) &= \left[\frac{\rho^2}{D^2} \right] \geq 0 && \text{for } d \geq n,\end{aligned}$$

where D is the minimal diameter of the sphere containing the set $(z_1, \dots, z_{\ell+k})$, are assigned to the d th element of the structure S_d . Utilizing the results of Theorem 8.4, one can show, as in Section 8.5, that the capacity of functions belonging to the S_d th element of the structure equals d , where

$$d = \min \left(\left[\frac{D^2}{\rho^2} \right] + 1, n \right).$$

The method of structural risk minimization for this structure involves finding an element S_* and a function $f(\mathbf{x}, \mathbf{a}; \cdot)$ in it such that the minimum on the right-hand side of the inequality

$$R_\Sigma(\alpha) \leq \left[\frac{1 + \tau a(p) \frac{\ell}{\ell+k} \mathcal{E}}{1 - \tau a(p) \frac{k}{\ell+k} \mathcal{E}} \right]_\infty R_{\text{emp}}(\alpha) \quad (8.49)$$

is obtained. Here \mathcal{E} is the smallest solution of the inequality

$$d \left(\ln \frac{\ell+k}{d} + 1 \right) + \ln \Gamma_{\ell,k}(\varepsilon) \leq \ln \eta.$$

The first factor on the right-hand side of (8.49) depends only on the order in which the vectors of the complete sample are projected on the vector of directions on the selected linear function, while the second factor depends on the value of the empirical risk.

Let the minimum on the right-hand side of (8.49) equal $R(\alpha_{\text{emp}}^*, d^*)$. Then the assertion

$$P\{R_\Sigma(\alpha_{\text{emp}}^*) < R(\alpha_{\text{emp}}^*, d^*)\} > 1 - n\eta.$$

is valid.

8.8 SAMPLE SELECTION FOR ESTIMATION OF VALUES OF REAL-VALUED FUNCTIONS

Let a complete sample

$$x_1, \dots, x_{\ell+k}, \quad (8.50)$$

be given. Consider $H'_{\ell+k}$ different subsets $X_1, \dots, X_{H'_{\ell+k}}$, each of which is obtained by omitting at most ℓ elements from (8.50). Below we shall assume that for all subsets the condition (8.16) is fulfilled.

Now let a training sequence

$$(y_1, x_1), \dots, (y_\ell, x_\ell) \quad (8.51)$$

and a working sequence

$$x_{\ell+1}, \dots, x_{\ell+k} \quad (8.52)$$

be defined on the initial set (8.50). The samples (8.51) and (8.52) induce on each of the subsets its own training and working samples, respectively.

Consider $H'_{\ell+k}$ problems of estimating values of a function at given points. For each problem r ($r = 1, \dots, H'_{\ell+k}$) we shall define—using the method described above—its own structure on the class of the class of linear functions

$$S'_1 \subset \dots \subset S'_n.$$

We then obtain with probability $1 - \eta$, for each of the problems (separately). the bound

$$R'_\Sigma(\alpha_{\text{emp}}) \leq \left[\frac{1 + \tau a(p) \frac{\ell - \ell'_{\text{exc}}}{\ell + k - \ell'_{\text{exc}} - k'_{\text{exc}}} \mathcal{E}^r}{1 - \tau a(p) \frac{k - k'_{\text{exc}}}{\ell + k - \ell'_{\text{exc}} - k'_{\text{exc}}} \mathcal{E}^r} \right]_\infty R'_{\text{emp}}(\alpha_{\text{emp}}) \quad (8.53)$$

is valid, where $f(x, \alpha_{\text{emp}}) \in S'_d$ is a function that minimizes the empirical risk on the training sequence for this problem (index r indicates that the overall and empirical risks are computed over the elements belonging to the subset X^r) and \mathcal{E}^r is the smallest solution of the inequality

$$d \left(\ln \frac{\ell + k - \ell'_{\text{exc}} - k'_{\text{exc}}}{d} + 1 \right) + \ln \Gamma_{\ell - \ell'_{\text{exc}}, k - k'_{\text{exc}}}(\mathcal{E}^r) \leq \ln \eta$$

Here we use the following notation: $\ell - \ell'_{\text{exc}}$ is the length of the training sequence in the problem r , and $k - k'_{\text{exc}}$ is the length of the working sample in the problem $\ell'_{\text{exc}} + k'_{\text{exc}} = t_r$. With probability $1 - \eta$ simultaneously for S_d elements of all $H'_{\ell+k}$ problems. the inequalities

$$R'_\Sigma(\alpha_{\text{emp}}) \leq \left[\frac{1 + \tau a(p) \frac{\ell - \ell'_{\text{exc}}}{\ell + k - \ell'_{\text{exc}} - k'_{\text{exc}}} \mathcal{E}^r}{1 - \tau a(p) \frac{k - k'_{\text{exc}}}{\ell + k - \ell'_{\text{exc}} - k'_{\text{exc}}} \mathcal{E}^r} \right]_\infty R'_{\text{emp}}(\alpha_{\text{emp}}) \quad (8.54)$$

hold true, where (unlike the preceding case) \mathcal{E}_r are the smallest solutions of the inequalities

$$d \left(\ln \frac{\ell + k - \ell_{\text{exc}}^r - k_{\text{exc}}^r}{d} + 1 \right) + \ln \Gamma_{\ell - \ell_{\text{exc}}^r, k - k_{\text{exc}}^r}(\varepsilon) + \ln H_{\ell+k}^r \leq \ln \eta.$$

We now choose a problem for which the bound of the value of the overall risk is minimal.

Finally, enumerating over d and t , we obtain the best solution.

8.9 LOCAL ALGORITHMS FOR ESTIMATING VALUES OF AN INDICATOR FUNCTION

Finally consider the local algorithms for estimating functional values. We define for each vector x_i of the complete sample a system of neighborhoods:

- (1) $(x_1)_1 \in (x_1, x_{i_1})_2 \in \dots \in (x_1, \dots, x_{\ell+k})_q$;
 - (2) $(x_2)_1 \in (x_2, x_{i_2})_2 \in \dots \in (x_1, \dots, x_{\ell+k})_q$
- $(\ell + k) \quad (x_{\ell+k})_1 \in (x_{\ell+k}, x_{i_{\ell+k}})_2 \in \dots \in (x_1, \dots, x_{\ell+k})_q \quad (8.55)$

Now let a subdivision of the set X into the training and the working sample be carried out.

Consider an arbitrary neighborhood X'_i of the point x , containing elements of both the training and the working samples. In view of Theorem 8.2, one can assert with probability $1 - \eta$ that simultaneously for all decision rules the inequality

$$\nu_{\Sigma}^r(\alpha) < \nu^r(\alpha) + \frac{k^r(\mathcal{E}^r)^2}{2(\ell^r + k^r)} + \varepsilon_*^r \sqrt{\nu^r(\alpha) + \left[\frac{k^r(\mathcal{E}^r)}{2(\ell^r + k^r)} \right]^2}$$

is fulfilled, where $\nu_{\Sigma}^r(\alpha)$ is the value of the overall risk of classification of elements belonging to the neighborhood X'_i by means of a decision rule $f(x, a)$, $\nu^r(\alpha)$ is the value of the empirical risk computed for the rule $f(x, a)$ based on the elements of the training sequence belonging to neighborhood X'_i , \mathcal{E}^r is the smallest solution of the inequality

$$n \left(\ln \frac{\ell^r + k^r}{n} + 1 \right) + \ln \Gamma_{\ell^r, k^r}(\varepsilon) \leq \ln \eta,$$

and n is the dimension of the space X . In this inequality, ℓ^r and k^r are the numbers of elements belonging to the neighborhood X'_i in the training and the working samples. Let $f(x, \alpha_{\text{emp}})$ be a decision rule that minimizes the value of the empirical risk on the training sequence belonging to X'_i .

For the elements belonging to X_i^r the bound

$$\nu_{\Sigma}^r(\alpha_{\text{emp}}) < \nu^r(\alpha_{\text{emp}}) + \frac{k^r(\mathcal{E}^r)^2}{2(\ell^r + k^r)} + \mathcal{E}^r \sqrt{\nu^r(\alpha_{\text{emp}}) + \left[\frac{k^r \mathcal{E}^r}{2(\ell^r + k^r)} \right]^2} = R_i(r)$$

is valid with probability $1 - \eta$. We shall now obtain a neighborhood of the point x_i for which the minimum (with respect to r) of the value $R_i(r)$ is attained. Let the minimum be attained in a neighborhood X_i^r and let $\omega_{i_1}, \dots, \omega_{i_s}$ be the classification of vectors obtained in the working sample belonging to this neighborhood. Clearly, with probability $1 - \eta q$ this classification contains less than $R_i(\tau)k_r = R_i$ errors.

Analogously, solutions can be obtained for neighborhoods of all vectors belonging to the population. The results are presented in Table 8.1.

In the first column of the table, the vectors are given which define the system of neighborhoods, followed by the best classification of vectors for the given system and finally the guaranteed bound on the number of classification errors. Observe that the same vectors of the working sample belong to the neighborhoods of different vectors and that the classifications of some vectors from the working sample presented in different rows of the second column may not be the same.

Denote by $\omega_{\ell+1}^*, \dots, \omega_{\ell+k}^*$ the correct classification of vectors from the working sample $x_{\ell+1}, \dots, x_{\ell+k}$. Then the content of the table may be written in the form

$$\begin{aligned} & \sum^{(1)} |\omega_i^* - \omega_i| < R_1 \\ & \dots \\ & \sum^{(\ell+k)} |\omega_i^* - \omega_i| < R_{\ell+k}, \end{aligned} \tag{8.56}$$

table 8.1. Results for neighborhoods of all vectors belonging to the population

Neighborhood of Point	Classification of Vectors					Bound on Overall Risk
	$x_{\ell+1}$...	$x_{\ell+j}$...	$x_{\ell+k}$	
x_1	ω_1^1	...	—	...	$\omega_{\ell+k}^1$	R_1
x_s	—	...	$\omega_{\ell+j}^s$...	—	R_s
$x_{\ell+k}$	—	...	—	...	$\omega_{\ell+k}^{\ell+k}$	$R_{\ell+k}$

where $\sum^{(x_i)}$ indicates that the summation is carried out only over those classifications of vectors of the working sample that belong to the selected neighborhood of the point x_i .

Each one of the inequalities (8.56) is fulfilled with probability $1 - q\eta$. Consequently the system is consistent (all inequalities are satisfied simultaneously) with probability exceeding $1 - q(\ell + k)\eta$.

Consider the set Ω of vectors $\bar{\omega} = (\bar{\omega}_{\ell+1}, \dots, \bar{\omega}_{\ell+k})$ of solution of the system (8.56). Actually the final vector of the classification may be chosen arbitrarily from this set. However, it is more expedient in such a case to choose a solution that possesses some additional extremal properties. Among all the vectors in Ω we shall find the minimax ω^* —that is, the one whose distance from the farthest vector belonging to the admissible set Ω is the smallest:

$$\omega^* = \operatorname{argmin}_{\omega \in \Omega} \max_{\bar{\omega} \in \Omega} |\omega - \bar{\omega}|.$$

The vector ω^* will be chosen as the final solution of the problem of classifying vectors in the working sample.

In this algorithm, by defining a system of neighborhoods of vectors in the complete sample, we were able to determine for each vector x_i an optimal neighborhood for constructing a linear decision rule. The rule thus obtained was used only for classification of the vectors to an optimal neighborhood. Such algorithms are sometimes referred to as *local* ones.

In practice, different ideas for defining neighborhood are utilized. In particular, a neighborhood X'_i of the vector x_i can be defined by means of metric closeness. (The set X'_i contains vectors belonging to the complete sample such that $\|x - x_i\| < c$, where c is a constant. The collection of constants $c_1 < \dots < c_l$ determines the system of neighborhoods.)

8.10 LOCAL ALGORITHMS FOR ESTIMATING VALUES OF A REAL-VALUED FUNCTION

Using the scheme described in the preceding section, one can construct local algorithms for estimating values of a real-valued function. Form a system of neighborhoods for vectors belonging to a complete sample:

$$(1) \quad (x_1)_1 \in (x_1, x_{i_1})_2 \in \dots \in (x_1, \dots, x_{\ell+k})_q$$

$$(\ell + k) \quad (x_{\ell+k})_1 \in (x_{\ell+k}, x_{j_1})_2 \in \dots \in (x_{i_1}, \dots, x_{\ell+k})_q$$

Let a subdivision of the set of vectors from the complete samples into elements belonging to training and working samples be carried out. Consider a system of neighborhood for the point x_i :

$$\begin{aligned} X_i^1 &\subset X_i^2 \dots \subset X_i^q, \\ X'_i &= (x_i, x_{i_2}, \dots, x_{i_p})_r. \end{aligned}$$

For each set X_i^r one can determine—using algorithms for estimating a linear function—the values of the function as well as a guaranteed bound on the value of the overall risk:

$$R_{\Sigma^{(r)}}(\alpha_{\text{emp}}) < \frac{1 + \tau a(p) \frac{\ell_r}{\ell_r + k_r} \mathcal{E}}{\left(1 - \tau a(p) \frac{k_r}{\ell_r + k_r} \mathcal{E}\right)_+} R_{\text{emp}}^{(r)}(\alpha_{\text{emp}}), \quad (8.57)$$

where \mathcal{E} is the smallest solution of the equation

$$n \left(\ln \frac{\ell_r + k_r}{n} + 1 \right) + \ln \Gamma_{\ell_r, k_r}(\varepsilon) \leq \ln \eta. \quad (8.58)$$

Here ℓ_r and k_r are the numbers of elements in the training sample and in the working sample belonging to X_i^r .

Choose the neighborhood of the point x_i and a function $F(x, \alpha_r^*)$ for which the bound (8.57) is minimal. Let k_r^* be the number of elements of the working sample belonging to this neighborhood. The inequality

$$\frac{1}{k_r^*} \sum_{k_r^*} (y_i - F(x_i, \alpha_{\text{emp}}^*))^2 \leq \mathcal{E}_r \quad (8.59)$$

is valid with probability $1 - q\eta$ for the value y_i belonging to this neighborhood obtained using the function $F(x, \alpha_{\text{emp}}^*)$. In (8.59) the summation is carried out over the vectors x from the working sample that belongs to the optimal neighborhood: y are the actual (unknown to us) values of the functional dependence at the points of the working sample, and $F(x_i, \alpha_{\text{emp}}^*)$ are the computed values. Thus for each point x_i (there are $\ell + k$ such points in total, which is the number of vectors in the complete sample) the inequality (8.59) is valid with probability $1 - \eta$. Therefore with probability $1 - q(\ell + k)\eta$ all $\ell + k$ inequalities

$$\begin{aligned} \frac{1}{k_1^*} \sum_{k_1^*} (y_i - F(x_i, \alpha_{\text{emp}}^*(1)))^2 &< \mathcal{E}_1 \\ \dots &\dots \\ \frac{1}{k_{\ell+k}^*} \sum_{k_{\ell+k}^*} (y_i - F(x_i, \alpha_{\text{emp}}^*(\ell + k)))^2 &< \mathcal{E}_{\ell+k} \end{aligned} \quad (8.60)$$

are fulfilled simultaneously.

Consider an admissible set $\{Y\}$ of solutions $(y_{\ell+1}, \dots, y_{\ell+k})$ of the system (8.60). This set is nonempty with probability $1 - q(\ell + k)\eta$. Choose as the response a solution Y^* such that its distance from the farthest point in $\{Y\}$

is the smallest (the minimax solution)—that is, a k -dimensional vector \mathbf{Y}^* for which the equality

$$\mathbf{Y}^* = \arg \min_{\mathbf{Y} \in \{\mathbf{Y}\}} \max_{\mathbf{Y} \in \{\mathbf{Y}\}} \rho(\mathbf{Y}, \mathbf{Y})$$

is valid.

8.11 THE PROBLEM OF FINDING THE BEST POINT IN A GIVEN SET

Consider the following problem. Given a training sample

$$(y_1, x_1), \dots, (y_\ell, x_\ell), \quad (8.61)$$

find among the working sample

$$x_{\ell+1}, \dots, x_{\ell+k} \quad (8.62)$$

the best point—that is, the point on which the unknown function takes the largest (or the smallest) value with the highest probability. Solving this problem we will try to avoid estimating the function at the points (8.62) (our goal is to find the *best point of the set* (8.62), not to estimate the values of the function). As before in similar situations it is possible that the available data (8.61) and (8.62) are insufficient for a reasonable solution of the intermediate problem (estimating the values of a function of the set (8.62)), but are sufficient for solving the desired problem (to find the best point of the set (8.62)).

Below we consider a particular solution to this problem, but first we would like to note that the statement of this problem is a response to the limited amount of available empirical data in some important settings of real-life problems.

Example. Only a few dozen antitumor drugs have been clinically tested on human beings to date. Meanwhile, hundreds of new antitumor drugs are synthesized annually. These drugs are tested using different models of human tumors (including animal tumors). Effectiveness of the drugs on the various models (on animals) does not, however, ensure its clinical effectiveness on humans. The problem is to identify in the given set of drugs (8.62) the clinically most active one, x_* . To identify such a drug, one can use the information x about the results of the models testing (8.62); and for the previous testing set of drugs, one can use both the information y about the clinical activity of the drug and the information x about activity on the models (pairs (8.61)).

Thus let the training sample (8.61) and the working sample (8.62) be given. Let the class of functions $f(x, a), a \in A$, contain a function $f(x, \alpha_0)$ that orders the vectors x from both the training and the working samples in the same way as an unknown function that determines the values of y . (For the

set of indicator functions, this condition degenerates into the requirement that $\alpha_0 \in A$.) Among the vectors of the working set, it is required to find the one x_* which with the highest probability possesses the maximal value $y_* = f(x_*, \alpha_0)$.

As before, first consider the case of indicator functions $f(x, \alpha), \alpha \in A$, where we denote $y = \omega \in \{0, 1\}$, and then we will consider the case where $f(x, a), a \in A$, is a set of real functions.

8.11.1 Choice of the Most Probable Representative of the First Class

Let $f(x, \alpha), \alpha \in A$, be a set of indicator functions. We denote

$$\begin{aligned} R &= (\omega_1, x_1), \dots, (\omega_\ell, x_\ell), \\ X_i &= x_{\ell+1}, \dots, \hat{x}_i, \dots, x_{\ell+k}, \\ \Omega_i &= \omega_{\ell+1}, \dots, \hat{\omega}_i, \dots, \omega_{\ell+k}. \end{aligned} \quad (8.63)$$

(The sequence X_i (the sequence Ω) is obtained from the sequence of all vectors (values) by omitting the element x_i (element ω_i).) Generally speaking, the sequence X_i can be divided into two classes in 2^{k-1} possible ways. Let us denote by Ω'_i , $r = 1, \dots, 2^{k-1}$, one of these ways. Assume that for each r the probability $P(\Omega'_i)$ is defined that Ω'_i will coincide with the classification of the sequence X_i performed using the function $f(x)$ that has been defined. Then for each fixed vector x_i of the working sample, one obtains

$$P\{\omega_i = 1|R, X\} = \sum_{r=1}^{2^{k-1}} P\{\omega_i = 1|R, X, \Omega'_r\} P(\Omega'_r). \quad (8.64)$$

Moreover, since the class $f(x, \alpha), \alpha \in A$, contains the function $f(x)$, one of the N equivalence classes F_1, \dots, F_N can separate data without an error. Let a priori probabilities for any equivalence class to perform without error be equal. The probabilities $P(\Omega'_r)$ and $P\{\omega_i = 1|R, X, \Omega'_r\}$ on the right-hand side (8.64) can be calculated. Namely, the probability that the classification Ω'_r of the vectors X_i coincides with that specified by the function $f(x)$ is equal to

$$P(\Omega'_r) = \frac{n(X_i, \Omega'_r)}{N}, \quad (8.65)$$

where $n(X_i, \Omega'_r)$ is the number of equivalence classes that classify the sequence in compliance with Ω'_r . The conditional probability that vector x ,

belongs to the class $\omega_i = 1$ is equal to

$$P\{\omega_i = 1 | R, X, \Omega_r^i\} = \chi_r^i = \begin{cases} 0 & \text{if there is no equivalence class that permits} \\ & \text{the separation } R \cup X_i \Omega_r^i; \\ \frac{1}{2} & \text{if there is an equivalence class that permits} \\ & \text{the separation } R \cup X_i \Omega_r^i \cup x_i, 1 \text{ and} \\ & \text{there is an equivalence class that permits} \\ & \text{the separation } R \cup X_i \Omega_r^i \cup x_i, 0; \\ 1 & \text{if there is an equivalence class that permits} \\ & \text{the separation } R \cup X_i \Omega_r^i \cup x_i, 1 \text{ and} \\ & \text{there is no equivalence class that permits} \\ & \text{the separation } R \cup X_i \Omega_r^i \cup x_i, 0. \end{cases} \quad (8.66)$$

Substituting the expression (8.65) and (8.66) into (8.64) we can estimate the probability

$$P\{\omega_i = 1 | R, X\} = \sum_{r=1}^{2^{k-1}} P\{\omega_i = 1 | R, X, \Omega_r^i\} \frac{n(X_i, \Omega_r^i)}{N}. \quad (8.67)$$

It remains to choose from the k vectors the one for which this probability is maximal.

Note that the larger the class $f(x, \alpha), \alpha \in A$, the smaller in general is

$$\max_i P\{\omega_i = 1 | R, X\}.$$

In the extreme case where the class of functions $f(x, \alpha), \alpha \in A$, is so wide that it permits the maximal possible number of equivalence classes, the equality

$$P\{\omega_i = 1 | R, X\} = \frac{1}{2}$$

holds no matter what the number i is.

Another natural assumption is that the a priori probability on the equivalence classes is given by binomial law with a parameter p (p is probability of occurrence of an element with $\omega = 1$). The a priori probability of correct separation is

$$p_j = \frac{C_{\ell+k}^{m_j} p^{m_j} (1-p)^{\ell+k-m_j}}{\sum_{j=1}^N C_{\ell+k}^{m_j} p^{m_j} (1-p)^{\ell+k-m_j}},$$

where m_j is the number of elements which are classified by the rules from F_j with $\omega = 1$. Under this assumption we have

$$P(\omega_j = 1 | R, X) = \sum_{r=1}^{2^{k-1}} \chi_r^j \frac{\sum_{j=1}^{n(X_i, \Omega_r^i)} C_{\ell+k}^{m_j} p^{m_j} (1-p)^{\ell+k-m_j}}{\sum_{j=1}^N C_{\ell+k}^{m_j} p^{m_j} (1-p)^{\ell+k-m_j}},$$

where

$$x_r^i = \begin{cases} 0 & \text{if there is no equivalence class that permits} \\ & \text{the separation } R \cup X_i \Omega_r^i; \\ \frac{P}{p + \frac{m_*^r(1-p)}{\ell + k + 1 - \frac{p}{m_*^r}}} & \text{if there is an equivalence class that permits} \\ & \text{the separation } R \cup X_i \Omega_r^i \cup x_i, 1 \text{ and there is an} \\ & \text{equivalence class that permits the separation} \\ & R \cup X_i \Omega_r^i \cup x_i, 0; \\ 1 & \text{if there is an equivalent class that permits} \\ & \text{the separation } R \cup X_i \Omega_r^i \cup x_i, 1 \text{ and there is no} \\ & \text{equivalence class that permits the separation} \\ & R \cup X_i \Omega_r^i \cup x_i, 0 \end{cases}$$

and m_*^r is the number of pairs in the set $R \cup X_i \Omega_r^i \cup x_i, 1$ with $\omega = 1$.

8.11.2 Choice of the Best Point of a Given Set

Consider the case where in a given training set

$$(y_1, x_1), \dots, (y_\ell, x_\ell)$$

y takes an arbitrary values. Note that the elements of the sequence

$$X = x_{\ell+1}, \dots, x_{\ell+k}$$

can be ordered in all possible ways by using permutation operators T_r ($r = 1, 2, \dots, k!$).

Let $T_0 X_0$ denote the sequence of the vectors x from the training set ordered according to the decreasing corresponding values y (for simplicity let us assume that the ordering is strict).

We write

$$z \succ^f X$$

if $f(z) > f(x)$ for all $x \in X$.

Assume now that the vector x_i is fixed in the working sample. There are $(k-1)!$ different ways to order the set X . Assume that for each of these one knows the probability $P(T_r X_i)$ that the order of the sequence $T_r X_i$ coincides with order of the sequence on X_i obtained by the decreasing corresponding elements $f(x)$. Then for each fixed vector x , of the working sample we obtain

$$P\{x_i \succ^f X_i | T_0 X_0, X\} = \sum_{r=1}^{(k-1)!} P\{x_i \succ^f X_i | T_0 X_0, T_r X_i\} P\{T_r X_i\}. \quad (8.68)$$

From the viewpoint of ordering the vector x of the complete sample, the class of functions $f(x, \alpha), \alpha \in A$, is decomposed into a finite number of equivalence classes F_1, \dots, F_N . Among these equivalence classes there is the one that

orders the complete sample set of x according to the decreasing values $f(x)$. Let there exist a priori probability defined on equivalence classes. Then the probabilities $P\{T_r X_i\}$ and $P\{x_i \succ^f X_i | T_0 X_0, T_r X_i\}$ can be computed. The probability that the ordering $T_r X_i$ coincides with the one given by the function $f(x)$ is

$$P\{T_r X_i\} = \frac{n(T_r X_i)}{N}, \quad (8.69)$$

where $n(T_r X_i)$ is the number of equivalence classes with the ordering $T_r X_i$. The conditional probability that x_i is the best point is

$$P\{x_i \succ^f X_i | T_0 X_0, T_r X_i\} = \chi_r^i = \begin{Bmatrix} 0 \\ \vdots \\ \frac{1}{p+1} \\ \vdots \\ 1 \end{Bmatrix}, \quad (8.70)$$

where χ_r^i is equal to zero if there is no equivalence class that permits simultaneously the ordering $T_0 X_0$ and $x_i, T_r X_i$ ($x_i, T_r X_i$ is a sequence whose leftmost element is x_i and the remaining ones coincide with $T_r X_i$) and χ_r^i is equal to $1/(p+1)$, $p = 0, \dots, k-1$, if there are $O, +1$ equivalence classes that permit simultaneously the ordering of $T_0 X_0$ and $x_i T_r X_i$. Substituting (8.69), (8.70), in (8.68), we obtain

$$P\{x_i \succ^f X_i, T_0 X_0, X\} = \sum_{r=1}^{(k-1)!} \frac{\chi_r^i \cdot n(T_r X_i)}{N} \quad (8.71)$$

Consequently, the probability that among the vectors of the working sample the vector x_i will have the maximum value off (x) is determined by (8.71). It remains to choose the vector with the maximum probability. As before, the wider the class of the functions $f(x, a), a \in A$, the smaller in general is

$$\max_i P\{x_i \succ^f X_i | T_0 X_0, X\}.$$

In the extreme case where the class of $f(x, a), a \in A$, is so wide that all $N = (\ell+k)!$ of equivalence classes are possible, the equality

$$P\{x_i \succ^f X_i | T_0 X_0, X\} = \frac{1}{k}$$

holds independent of the number of i .



SUPPORT VECTOR ESTIMATION OF FUNCTIONS

Part II introduces methods that provide generalization when estimating multi-dimensional functions from a limited collection of data.

9

PERCEPTRONS AND THEIR GENERALIZATIONS

The next chapters of the book are devoted to constructing algorithms for pattern recognition, regression estimation, signal processing problems, and solving linear operator equations. Methods for constructing hyperplanes with good statistical properties are the key to these algorithms.

In this chapter we consider the simplest methods. These methods bring us to the classical learning algorithms: perceptrons and their generalizations, potential functions (radial basis functions), and neural networks.

The basis for these methods were developed in the 1960s. In the late 1980s and early 1990s they became the major tools for constructing learning machines.

9.1 ROSENBLATT'S PERCEPTRON

At the end of the 1950s, F. Rosenblatt proposed a machine (Perceptron) for learning from examples. He considered a set of indicator functions[†] linear in their parameters

$$f(x, w) = \text{sign} \left\{ \sum_{p=1}^n w^p \psi_p(x) \right\} \quad (9.1)$$

and suggested a procedure for choosing from this set an approximating function using a sequence of examples

$$(y_1, x_1), \dots, (y_\ell, x_\ell),$$

[†]To simplify the notations we consider indicator functions of the form $\text{sign}(u) \in \{-1, 1\}$ instead of the form $\theta(u) \in \{0, 1\}$.

where $y_i = 1$ if vector x_i belong to the first class and $y_i = -1$ if vector x_i does not belong to the first class.

The Perceptron utilizes the following recurrent procedure for choosing the function (the coefficients $w = (w^1, \dots, w^n)$ in Eq. (9.1)):

1. At step zero choose the function $f(x, 0)$ (with coefficients $w(0) = (0, \dots, 0)$).
2. At step t using the element (y_t, x_t) of the training sequence, change the vector of coefficients $w(t - 1)$ in accordance with the rule

$$w(t) = \begin{cases} w(t - 1) & \text{if } y_t(w(t - 1) * \Psi_t) > 0, \\ w(t - 1) + y_t \Psi_t & \text{if } y_t(w(t - 1) * \Psi_t) \leq 0, \end{cases}$$

where we denoted by $\Psi_t = (\psi_1(x_t), \dots, \psi_n(x_t))$ an n -dimensional vector and by $(w(t - 1) * \Psi_t)$ the inner product of two vectors.

Note that the coefficients $w(t - 1)$ change only if the example (y_t, x_t) is misclassified by the constructing hyperplane.

In Rosenblatt's Perceptron the functions $\psi_1(x), \dots, \psi_n(x)$ are superpositions of fixed linear indicator functions. However, for the theory of Perceptrons this fact is not important. Without restriction of generality we will assume below that there exists a nonlinear operator, A , mapping vectors $x \in X$ into vector $u \in U$. Therefore the Perceptron constructs a separating hyperplane

$$f(u, w) = \text{sign}\{(u * w)\}$$

passing through the origin in U space. The space U was called *feature space*. The problem of constructing the nonlinear decision rule in the space X reduces to constructing a separating hyperplane in the space U . In this space the rule for estimating unknown parameters has the form

$$w(t) = \begin{cases} w(t - 1), & \text{if } y_t(w(t - 1) * u_t) > 0, \\ w(t - 1) + y_t u_t & \text{if } y_t(w(t - 1) * u) \leq 0. \end{cases} \quad (9.2)$$

In the early 1960s the first theorems concerning the Perceptron's algorithm were proved and these theorems actually started learning theory. In this section we discuss two of these theorems (the proofs of the theorems are given in the next section).

Consider an infinite sequence of examples (in feature space)

$$\{\hat{Y}, \hat{U}\} = (y_1, u_1), \dots, (y_\ell, u_\ell), \dots \quad (9.3)$$

Suppose that there exists such a vector, w_0 , that for some $\rho_0 > 0$ the inequality

$$\min_{(y, u) \in \{\hat{Y}, \hat{U}\}} \frac{y(w_0 * u)}{|w_0|} \geq \rho_0 \quad (9.4)$$

holds true. (The given examples are separable with margin ρ_0 .) Then the following theorem is true.

Theorem 9.1 (Novikoff). *Let the infinite sequence of training examples (9.3) with elements satisfying the inequality*

$$|u_i| < D \quad (9.5)$$

be given. Suppose that there exists a hyperplane with coefficients w_0 that separates correctly elements of this sequence and satisfies the condition (9.4).

Then using the iterative procedure (9.2), the Perceptron constructs a hyperplane that correctly separates all examples of this infinite sequence. To construct such a hyperplane the Perceptron makes at most

$$M = \left\lceil \frac{D^2}{\rho_0^2} \right\rceil$$

corrections, where $[a]$ is the integer part of the value a .

Note that this theorem does not take into account how the sequence was chosen. To make some statistical inference from the Novikoff theorem we consider two cases:

Case 1. An infinite sequence of examples (9.3) is a training sequence[†] of size ℓ repeated an infinite number of times.

Case 2. An infinite sequence (9.3) containing pairs which are chosen randomly and independently in accordance with some fixed distribution.

Case 1: An infinite sequence is a training sequence of size ℓ repeated an infinite number of times. In this case the theorem asserts that under conditions (9.4) the proposed procedure minimizes up to zero the functional of empirical risk and that it will be done after a finite number M of corrections. In this case, using the bound obtained for the empirical risk minimization principle, one can get a bound on the risk for a decision rule constructed by the Perceptron: With probability at least $1 - \eta$ the inequality

$$R(w_\ell) \leq \frac{n \left(\ln \frac{2\ell}{n} + 1 \right) - \ln \eta/4}{\ell} \quad (9.6)$$

holds true. (Recall that the VC dimension of the Perceptron with n weights is equal to n and the empirical risk after M corrections equal to zero.)

Also the following theorem is true.

[†]Recall that a training sequence is a sequence of pairs drawn randomly and independently in accordance with a fixed but unknown distribution.

Theorem 9.2. *For the Perceptron algorithm of constructing hyperplanes in the regime that separates the training data without error, the following bound on error rate is valid*

$$ER(w_\ell) \leq \frac{E \left[\frac{D_\ell^2}{\rho_\ell^2} \right]}{\ell + 1}$$

where the expectation is taken over the random value $\left[\frac{D_\ell^2}{\rho_\ell^2} \right]$ (defined by the values D_ℓ and ρ_ℓ calculated from the training data).

The proof of this theorem is given in Section 10.4.4.

Case 2: An infinite sequence of examples chosen randomly and independently. In this case under conditions of Theorem 9.1 after M corrections there is an infinite tail of the sequence that will be separated without error.

However, Theorem 9.1 does not tell us how many examples should be considered to make all necessary corrections. On the other hand we are looking for a solution with a small value of risk, $\varepsilon > 0$ (probability of errors on the tail of the sequence), which is not necessarily equal to zero. In this situation (as we will show below) using information about the value of the margin ρ_0 in (9.4) and value of the bound D in (9.5), one can estimate the size of the training subsequence that one needs to guarantee the construction of a hyperplane that with given probability $1 - \eta$ possesses probability of error at most ε .

To define the size of such a subsequence let us introduce the idea of **stopping rules** for the learning processes. Let us simultaneously do two things: Construct a hyperplane and evaluate its quality. If this quality is not high, we will continue the learning process; otherwise we will stop it.

We use the following rule: Stop the learning process if after k corrections the next $m(k)$ elements of the training sequence do not change the obtained hyperplane (next $m(k)$ examples are recognized correctly).

The theory of stopping rules should answer two questions:

1. What the values $m(k)$ should be to guarantee that if the algorithm will stop, then the constructed decision rule with high probability $1 - \eta$ has the risk at most ε .
2. On what size of training sequence the learning process will be stopped.

The following theorem and its corollary provide the answers to these questions.

Theorem 9.3. *Let the learning process be stopped in accordance with the stopping rule. Then with probability $1 - \eta$, one can assert that the constructed de-*

cision rule has a risk at most ε if the values $m(k)$ are defined by the following equalities:

$$m(k) = \left\lceil \frac{2 \ln k - \ln \eta + \ln \frac{\pi^2}{6}}{-\ln(1-\varepsilon)} \right\rceil + 1, \quad k = 1, 2, \dots \quad (9.7)$$

Corollary. Let, during the learning process, the Perceptron count how many examples of the training sequence do not affect the decision rule after k corrections and stop the learning process if for some k this number is equal to $m(k)$, as given by Eq. (9.7).

Then under the condition of Theorem 9.1, one can assert that with probability 1 the Perceptron will stop the learning process at most at the $(\ell + 1)$ st example, where

$$\ell = \left\lceil 1 + \frac{2 \ln \left[\frac{D^2}{\rho_0^2} \right] - \ln \eta + \ln \frac{\pi^2}{6}}{-\ln(1-\varepsilon)} \right\rceil \left[\frac{D^2}{\rho_0^2} \right]. \quad (9.8)$$

To prove the corollary, note that according to Theorem 9.1 the number of corrections does not exceed $\left[\frac{D^2}{\rho_0^2} \right]$ and according to Theorem 9.2 the largest interval between corrections is $m \left(\left[\frac{D^2}{\rho_0^2} \right] \right)$. So the learning process will stop on the part of the sequence of size less than

$$\ell = \left[\frac{D^2}{\rho_0^2} \right] \times m \left(\left[\frac{D^2}{\rho_0^2} \right] \right).$$

Equation (9.8) is explicit form of this equation.

Remark. Note that the corollary does not assert that after ℓ steps the risk for the constructed hyperplane will not exceed ε . This theorem guarantees that during the first ℓ steps the learning process will stop and that in this moment the constructed hyperplane will possess the required quality. If at this moment the algorithm does not stop, then the next iterations can worsen the quality of the hyperplane and after ℓ steps this quality can be worse than the required one.

Thus we considered two different regimes of learning for the Perceptron. In the first regime the Perceptron used the same training sequence several

times to separate training examples. For this regime we can assert that with probability $1 - \eta$ the obtained risk is bounded by inequality (9.6).

In the second regime the Perceptron uses any element of the training sequence only one time and stops the learning process in accordance with the stopping rule (which depends on ε). For this regime, one can assert that if ε in Eq. (9.8) is such that

$$\varepsilon \geq 1 - \exp \left\{ - \frac{2 \ln \left[\frac{D^2}{\rho_0^2} \right] - \ln \eta + \ln \frac{\pi^2}{6}}{\ell} \left[\frac{D^2}{\rho_0^2} \right] \right\}, \quad (9.9)$$

then with probability 1 the learning process will be stopped on the sequence of size ℓ , bounded by equality (9.8) and with probability $1 - \eta$ that the risk for the chosen function will be less than ε . (This assertion is a consequence of Theorem 9.3 and Eq. (9.9).)

9.2 PROOFS OF THE THEOREMS

9.2.1 Proof of Novikoff Theorem

Consider a training sequence

$$(y_1, u_1), \dots, (y_\ell, u_\ell),$$

where y_1 is equal to 1 if vector u_i belongs to the first class and equal to -1 if this vector belongs to the second class. Then according to the Perceptron's algorithm:

1. If on iteration t the vector u_t is recognized correctly that is,

$$y_t(u_t * w(t-1)) > 0,$$

then vector $w(t)$ is not changed, that is,

$$w(t) = w(t-1).$$

2. If this vector misclassified, that is,

$$y_t(u_t * w(t-1)) \leq 0, \quad (9.10)$$

then vector $w(t)$ is changed according to

$$w(t) = w(t-1) + y_t u_t.$$

We say that in this case the vector of coefficients was corrected.

3. The initial vector $w(0) = 0$.

Let us bound the norm of vector $w(k)$ after k corrections. Note that if on step t the correction was done, then

$$|w(k)|^2 = |w(k-1)|^2 + 2y_t(u_t * w(k-1)) + |u_t|^2$$

Taking into account that in this case the inequality (9.10) is valid and

$$|u_t| \leq D,$$

one obtains

$$|w(k)|^2 \leq |w(k-1)|^2 + D^2$$

Therefore if at step t the number of corrections is equal to k , then

$$|w(k)|^2 \leq kD^2 \quad (9.11)$$

since $w(0) = 0$.

Now in accordance with the condition of the theorem there exists a vector w_0 such that the inequality

$$\frac{y_i(u_i * w_0)}{\|w_0\|} \geq p_0$$

holds true. Let us bound the inner product in $(w(k) * w_0)/\|w_0\|$. If at the step t the k th correction occurs, then the inequality

$$\begin{aligned} \frac{(w(k) * w_0)}{\|w_0\|} &= \frac{(w(k-1) * w_0)}{\|w_0\|} + \frac{y_t(u_t * w_0)}{\|w_0\|} \\ &\geq \frac{(w(t-1) * w_0)}{\|w_0\|} + p_0 \end{aligned}$$

is valid.

Therefore, the inequality

$$\frac{(w(k) * w_0)}{\|w_0\|} \geq kp_0 \quad (9.12)$$

is valid. Using the Cauchy–Schwartz inequality

$$(w(k) * w_0) \leq \|w(k)\| \|w_0\|$$

we have

$$\|w(k)\| \geq kp_0. \quad (9.13)$$

Combining inequalities (9.11) and (9.13), we obtain

$$k \leq \frac{D^2}{p_0^2}.$$

Therefore the number of corrections does not exceed $[D^2/p_0^2]$.

The theorem has been proved.

9.2.2 Proof of Theorem 9.3

During the learning process, let the decision rules

$$F(x, w_1), \dots, F(x, w_k), \dots$$

be chosen. Now estimate the probability that the learning process will be stopped in the moment when the chosen rule $F(x, w_k)$ has a probability of error $P(w_k) > \varepsilon$. The probability P_k that the learning process will be stopped after k corrections, but before the $(k+1)$ th correction, is equal to the probability of the event that after k corrections occur, $m(k)$ correct classifications are made by decision rule $F(x, w_k)$ which is given by

$$P_k = (1 - P(w_k))^{m(k)} < (1 - \varepsilon)^{m(k)}.$$

Therefore the probability that the learning process will be stopped when $P(w_k) > \varepsilon$ can be bounded as

$$P < \sum_{k=1}^{\infty} P_k < \sum_{k=1}^{\infty} (1 - \varepsilon)^{m(k)}.$$

Let us choose the integer-valued function $m(k)$ such that

$$(1 - \varepsilon)^{m(k)} \leq \frac{a}{k^n}, \quad n > 1. \quad (9.14)$$

From this equality, one can find that

$$m(k) \leq \left\lceil \frac{\ln a - n \ln k}{\ln(1 - \varepsilon)} \right\rceil + 1 \quad (9.15)$$

It remains to determine the constant a in such a way that inequality

$$P < \sum_{k=1}^{\infty} (1 - \varepsilon)^{m(k)} \leq \eta$$

is valid. To do this we use the inequality (9.14). We obtain

$$\sum_{k=1}^{\infty} \frac{a}{k^n} = a \zeta(n) = \eta,$$

where

$$\zeta(n) = \sum_{k=1}^{\infty} \frac{1}{k^n}, \quad n > 1.$$

From this equality, one can find

$$\ln a = \ln \eta - \ln \zeta(n). \quad (9.16)$$

Therefore from (9.15) and (9.16) one can find that under conditions

$$m(k) = \left[\frac{\ln q - \ln \zeta(n) - n \ln k}{\ln(1 - \varepsilon)} \right] + \dots \quad (9.17)$$

the probability P does not exceed the required value η . The bound (9.17) is valid for any $n > 1$. In particular for $n = 2$ we have $\zeta(2) = \pi^2/6$. Therefore

$$m(k) = \left[\frac{-\ln \eta + 2 \ln k + \ln \frac{\pi^2}{6}}{-\ln(1 - \varepsilon)} \right] + 1.$$

The theorem has been proved.

9.3 METHOD OF STOCHASTIC APPROXIMATION AND SIGMOID APPROXIMATION OF INDICATOR FUNCTIONS

Previous sections considered the problem of minimizing the risk in the set of linear indicator functions

$$f(u, w) = \text{sign}\{(u * w)\} \quad (9.18)$$

under conditions that a given training sequence can be separated by a linear hyperplane—that is, when there exists $\rho_0 > 0$ such that the inequality

$$\min_{(y,u) \in \{\hat{Y}, \hat{U}\}} \frac{y(w_0 * u)}{|w_0|} \geq \rho_0 \quad (9.19)$$

holds true. In particular we considered the case of minimizing the empirical risk functional (see Section 9.1, Case 1). Now, we would like to construct a hyperplane that minimizes the risk when the training set

$$(y_1, u_1), \dots, (y_\ell, u_\ell)$$

cannot be separated by a hyperplane without error. That is, there is no w_0 that satisfy the inequality (9.19) for a given small ρ_0 .

It is known that the problem of minimizing the number of training errors is NP complete. Therefore one tries to find the vector w_0 that provides the local minimum to the risk functional

$$R(w) = \int (y - \text{sign}\{(u * w)\})^2 dP(u, y) \quad (9.20)$$

or the local minimum to the empirical risk functional

$$R_{\text{emp}}(w) = \frac{1}{\ell} \sum_{j=1}^{\ell} (y_j - \text{sign}\{(u_j * w)\})^2. \quad (9.21)$$

9.3.1 Method of Stochastic Approximation

In the 1960s a general method for minimizing risk functionals, the so-called stochastic approximation method, was discovered. According to this method in order to minimize the risk functional

$$R(\alpha) = \int Q(z, \alpha) dP(z) \quad (9.22)$$

using the i.i.d. data

$$z_1, \dots, z_\ell, \dots, \quad (9.23)$$

one has to define the gradient $\text{grad}_\alpha Q(z_i, \alpha)$ (with respect to α for given z) of the function $Q(z, \alpha)$.

Suppose that at any point z , one can estimate the gradient with some independent random noise

$$\text{grad}_\alpha Q(z_i, \alpha) + \xi_i.$$

To find a local minimum of the functional (9.22) using data (9.23), the stochastic approximation method suggests the following recurrent procedure:

$$\alpha_t = \alpha_{t-1} - \gamma_t [\text{grad}_\alpha Q(z_t, \alpha_{t-1}) + \xi_t], \quad (9.24)$$

where sequence of values $\gamma_t \geq 0$ satisfies the conditions

$$\begin{aligned} \lim_{t \rightarrow \infty} \gamma &= 0 \\ \sum_{t=1}^{\infty} \gamma_t &= \infty, \\ \sum_{t=1}^{\infty} \gamma_t^2 &< \infty, \end{aligned} \quad (9.25)$$

and

$$E(\xi | \alpha) = 0. \quad (9.26)$$

It was proven that this method is consistent under very general conditions. Below we describe one version of conditions of these type.

Theorem 9.4 (Litvakov). *Let the following conditions hold:*

1. *The functional $R(\alpha)$, $\alpha \in A$, is bounded from below and there exists a nonempty set of local minima*

$$T = \{ \alpha : R(\alpha) = \inf R(\alpha) \}.$$

2. *The norm of gradients is bounded as follows:*

$$E|\text{grad}_\alpha Q(z, \alpha)|^2 < D(1 + |\alpha|^2). \quad (9.27)$$

3. *The random noise satisfies the following conditions:*

$$\begin{aligned} E(\xi|\alpha) &= 0, \\ E(\xi^2|\alpha) &< D(1 + |\alpha|^2). \end{aligned} \quad (9.28)$$

Then for any initial point α_ with probability 1 the convergence*

$$R(\alpha_i) \xrightarrow[i \rightarrow \infty]{} \inf R(\alpha)$$

is valid.

It was also proved that under conditions of Theorem 9.4 the stochastic approximation procedure (9.24) converges to a (local) minimum of the empirical risk functional

$$R_{\text{emp}}(\alpha) = \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i \alpha), \quad (9.29)$$

where sequence (9.23) is a sequence z_1, \dots, z_ℓ repeated an infinite number of times.

9.3.2 Sigmoid Approximations of Indicator Functions

One cannot apply stochastic approximation procedures for reaching a local minimum of functional (9.20), since for this functional the gradient of the loss function

$$Q(z, w) = (y - \text{sign}\{(y - (u^* w))\})^2$$

does not satisfy the conditions for consistency of the stochastic approximation procedure (the gradient of function $\text{sign}\{(y - (u^* w))\}$ is defined by the delta function).

Therefore, the idea was proposed to approximate the indicator functions $\text{sign}\{(y - (w^* u))\}$ by the so-called *sigmoid functions* (Fig. 9.1)

$$\tilde{f}(u, w) = S\{(w^* u)\}, \quad (9.30)$$

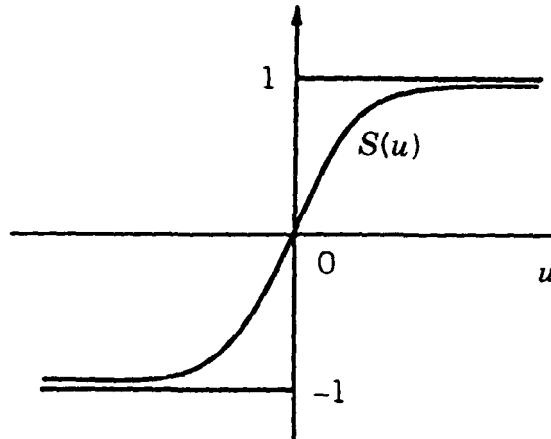


FIGURE 9.1. The sigmoid approximation of the sign function.

where $S(u)$ is a smooth monotonic function such that

$$S(-\infty) = -1, \quad S(+\infty) = 1,$$

for example,

$$S(a) = \tanh a = \frac{\exp(a) - \exp(-a)}{\exp(a) + \exp(-a)}.$$

For sigmoid functions, the gradient

$$\text{grad} [y - S((w * u))]^2 = -2[y - S((w * u))]S'_w\{(w * u)\}u$$

satisfies the conditions described in Theorem 9.3 and therefore for sigmoid approximation of indicator functions one can use a stochastic approximation procedure (9.24). One can also use a gradient descent procedure for minimizing the empirical risk functional with sigmoid approximation of the indicator function

$$R_{\text{emp}}(w) = \frac{1}{\ell} \sum_{j=1}^{\ell} (y_j - S\{(w * u_j)\})^2.$$

Using the gradient of the empirical risk functional

$$\text{grad}_w R_{\text{emp}}(w) = -\frac{2}{\ell} \sum_{j=1}^{\ell} [y_j - S((w * u_j))] S'\{(w * u_j)\} u_j,$$

one defines the procedure

$$w_t = w_{t-1} - \gamma_t \text{grad} R_{\text{emp}}(w_{t-1}),$$

where $\gamma_t > 0$ are values that satisfy the conditions (9.25).

For convergence of the gradient descent method to a local minimum, it is sufficient that the conditions of Theorem 9.3 be satisfied. Thus, the idea is to use the sigmoid approximation at the stage of estimating the coefficients and to use the indicator functions (with the obtained coefficients) at the stage of recognition.

The idea of sigmoid approximation of indicator functions was proposed in the mid-1960s. However, it became popular in the mid-1980s when the method of evaluating the gradient for an entire set of neurons forming the Multilayer Perceptron (the so-called back-propagation method) was used. We will discuss this method in Section 9.6; however, first we consider another idea of the 1960s, which in the middle of the 1980s was used for creating another approach to the problems of function estimation, the so-called Radial Basis Function approach.

9.4 METHOD OF POTENTIAL FUNCTIONS AND RADIAL BASIS FUNCTIONS

In the mid-1960s Aizerman, Braverman, and Rozonoer (1964a,b) suggested the so-called method of potential functions, where they proposed to estimate the functional dependency from the data

$$(y_1, x_1), \dots, (y_\ell, x_\ell)$$

using the following set of functions:

$$f(x, \alpha) = \text{sign} \left\{ \sum_{i=1}^{\ell} \alpha_i \phi(|x - x_i|) \right\} \alpha, \quad (9.31)$$

where $\psi(0) = 1$ and $\lim_{u \rightarrow \infty} \phi(|u|) = 0$. Function $\phi(|u|)$, called a potential function (by analogy with physical potentials), is a monotonic function that converges to zero with increasing $|u|$. For example,

$$\phi(|u|) = \exp\{-\gamma|u|\}.$$

Using function $f(x, \alpha)$, vector x^* is classified by the following rule: Vector x^* belongs to the first class if $f(x^*, \alpha) > 0$; otherwise, it belongs to the second class.

The algorithms proposed for the method of potential functions were essentially on-line algorithms — that is, algorithms that in order to construct an approximation use at any moment only one element of the training data (like the Perceptron algorithm; see Section 9.1). By 1970 the theory of consistency of the method of potential functions was constructed.

After 20 years the interest in this type of functions appeared again. In the mid-1980s, functions of this type were called radial basis functions (RBFs).

However, this time, to construct radial basis function approximations to the desired function, the off-line methods, namely the methods that minimize the empirical risk, were used.

9.4.1 Method of Potential Functions in Asymptotic Learning Theory

When solving the pattern recognition problem on the basis of the potential functions method, one distinguishes between two cases: the case where representatives of two different classes can be separated by the potential function (9.31) (the so-called deterministic setting of the pattern recognition problem) and the case where representatives of two different classes cannot be separated by the potential function (9.31) (the so-called stochastic setting of the pattern recognition problem).

For the deterministic case the Perceptron algorithm is used. Let the approximation

$$f_{t-1}(x) = \sum_{i=1}^t \alpha_i \phi(|x - x_i|)$$

to the desired function be constructed by the step number t . Then on the t th step the algorithm makes the following correction based on a new element of the training set (x_t, y_t) :

$$f_t(x) = \begin{cases} f_{t-1}(x) & \text{if } y_t f_{t-1}(x_t) > 0, \\ f_{t-1}(x) + y_t \phi(|x - x_t|) & \text{if } y_t f_{t-1}(x_t) \leq 0. \end{cases}$$

The key idea for proving the convergence of the method of potential functions is that for any chosen potential function $\phi(x)$ satisfying some constraints^t there exists a feature space Z (not necessarily finite)

$$z_1 = \xi(x), \dots, z_N = \xi_N(x), \dots,$$

where the potential function (9.31) has an equivalent representation as a separating hyperplane. Therefore, the theorems described in Section 9.1 for the Perceptron can be applied for the case of potential functions. Using the stopping rule described in Section 9.1, one can prove that for the deterministic case the algorithm constructs the desired approximation in a finite number of steps.

For the stochastic setting of the learning problems the method of potential functions relies on the stochastic approximation method of constructing approximations described in Section 9.3:

$$f_t(x) = f_{t-1}(x) + 2\gamma_t [y_t - f_{t-1}(x_t)] \phi(|x - x_t|), \quad (9.32)$$

^t We discuss these constraints in Chapter 10, Section 10.8.4.

where $\gamma_t > 0$ are values that satisfy the general rule for the stochastic approximation processes

$$\begin{aligned} \lim_{t \rightarrow \infty} \gamma_t &= 0 \\ \sum_{t=1}^{\infty} \gamma_t &= \infty, \\ \sum_{t=1}^{\infty} \gamma_t^2 &< \infty. \end{aligned} \quad (9.33)$$

The procedure (9.32) was also used to solve the regression problem. It was also shown that if y takes two values, namely zero and one, then for the pattern recognition case the procedure (9.32) led to estimating the conditional probability function—that is, estimating the probability that vector \mathbf{x} belongs to the first class.

9.4.2 Radial Basis Function Method

In the middle of the 1980s the interest in approximations using a set of potential functions (radial basis functions) reappeared. However, this time instead of the stochastic approximation inference—the ERM method for constructing the approximation was used. In other words, to construct the approximation in the set of radial basis functions (RBF) one minimizes the empirical risk functional

$$R_{\text{emp}}(\alpha) = \sum_{i=1}^{\ell} \left(y_i - \sum_{j=1}^{\ell} \alpha_j \phi(|x_i - x_j|) \right)^2. \quad (9.34)$$

The conditions were found (see Section 10.8.4) under which the matrix $A = ||a_{ij}||$ with the elements $a_{ij} = \phi(|x_i - x_j|)$ is positive definite and therefore the problem of minimizing (9.34) has a unique solution.

Later the RBF method was modernized where kernels were defined not at every point of the training data but at some specific points c_j , $j = 1, \dots, N$ (called centers):

$$R_{\text{emp}}(\alpha) = \sum_{i=1}^{\ell} \left(y_i - \sum_{j=1}^N \alpha_j \phi(|x_i - c_j|) \right)^2. \quad (9.35)$$

Several heuristics (mostly based on nonsupervised learning procedures) for specifying both the number N of centers and the positions c_j , $j = 1, \dots, N$, of the centers were suggested. In Chapter 10 we will define these elements automatically using new learning techniques based on the idea of constructing a specific (optimal) separating hyperplane in the feature space.

9.5 THREE THEOREMS OF OPTIMIZATION THEORY

Before we continue description of the learning algorithms, let us describe the main tools of optimization that we will use for constructing learning algorithms in this and the next chapters.

9.5.1 Fermat's Theorem (1629)

The first general analytical method for solving optimization problems was discovered by Fermat in 1629. He described a method for finding the minimum or maximum of functions defined in entire space (without constraints):

$$f(x) \rightarrow \text{extr.}$$

Let us start with the one-dimensional case.

A function $f(x)$ defined on R^1 is called differential at the point x^* if there exists α such that

$$f(x^* + A) = f(x^*) + a\lambda + r(\lambda),$$

where $r(\lambda) = o(|\lambda|)$; that is, for any small $\varepsilon > 0$ there exists $\delta > 0$ such that the condition

$$|\lambda| < \delta$$

implies the inequality

$$|r(\lambda)| < \varepsilon |\lambda|.$$

The value α is called the *differential off at point x^** and is denoted $f'(x^*)$. Therefore

$$f'(x^*) = \lim_{A \rightarrow 0} \frac{f(x^* + \lambda) - f(x^*)}{A} = a.$$

Theorem (Fermat). *Let $f(x)$ be a function of one variable, differentiable at point x^* . If x^* is a point of local extremum, then*

$$f'(x^*) = 0. \quad (9.36)$$

The point x^ for which (9.36) holds is called the stationary point.*

A function $f(x)$ defined on R^n is called differentiable at point $x^ = (x_1^*, \dots, x_n^*)$ if there exist values $a = (\alpha_1, \dots, \alpha_n)$ such that*

$$f(x^* + h) = f(x^*) + \sum_{i=1}^n \alpha_i h_i + r(h),$$

where $r(h) = o(|h|)$; that is, for any $\varepsilon > 0$ there exists $\delta > 0$ such that

$$|h| = \sqrt{h_1^2 + \dots + h_n^2} < \delta,$$

which implies

$$|r(h)| \leq \varepsilon h.$$

The collection $\mathbf{a} = (a_1, \dots, a_n)$ is called the derivative of the *function* $f(x)$ at the point x^* and is denoted $f'(x^*)$. Note that $f'(x^*)$ is a collection of n values. The value a_i ,

$$a_i = \lim_{\lambda \rightarrow 0} \frac{f(x^* + \lambda e_i) - f(x^*)}{\lambda}$$

where $e_i = (0, \dots, 1, \dots, 0)$ is called the i th partial derivative and it is denoted by $f'_{x_i}(x^*)$ or $\partial f(x^*) / \partial x_i$. Therefore $f'(x^*) = (f'_{x_1}(x^*), \dots, f'_{x_n}(x^*))$.

Corollary (Fermat's theorem for function of n variables). Let f be a *function* of n variables differentiable at point x^* . If x^* is a point of local *extremum of* the function $f(x)$, then

$$f'(x^*) = 0;$$

that is,

$$f'_{x_1}(x^*) = \dots = f'_{x_n}(x^*) = 0. \quad (9.37)$$

Fermat's theorem shows a way to find the stationary points of functions (that satisfy the necessary conditions to be a minimum or a maximum point). To find these points it is necessary to solve a system (9.37) of n equations with n unknown values $x^* = (x_1^*, \dots, x_n^*)$.

9.5.2 Lagrange Multipliers Rule (1788)

The next important step in optimization theory was done more than 150 years later when Lagrange suggested his rule for solving the following so-called conditional optimization problem: Minimize (or maximize) the function (of n variables)

$$f_0(x) \rightarrow \min \quad (9.38)$$

under constraints of equality type

$$f_1(x) = \dots = f_m(x) = 0. \quad (9.39)$$

Here we consider functions $f_r(x)$, $r = 0, 1, \dots, m$, that possess some smoothness (differentiability) properties. We assume that in subset X of the space R^n all functions $f_r(x)$, $r = 0, 1, \dots, m$, and their partial derivatives are continuous.

We say that $x^* \in X$ is a point of local minimum (maximum) in the problem of minimizing (9.38) under constraint (9.39) if there exists $\varepsilon > 0$ such that for any x that satisfy conditions (9.39) and constraint

$$|x - x^*| < \varepsilon$$

the inequality

$$f_0(x) \geq f_0(x^*)$$

(or the inequality

$$f_0(x) \leq f_0(x^*)$$

holds true.

Consider the function

$$L(x, \lambda, \lambda_0) = \sum_{k=0}^m \lambda_k f_k(x) \quad (9.40)$$

called the **Lagrange** function or Lagrangian and also consider the values $\lambda_0, \lambda_1, \dots, \lambda_m$ called **Lagrange** multipliers.

Theorem (Lagrange). *Let the functions $f_k(x)$, $k = 0, 1, \dots, m$, be continuous and differentiable in a vicinity of point x^* . If x^* is the point of a local extremum, then one can find Lagrange multipliers $\lambda^* = (\lambda_0^*, \dots, \lambda_m^*)$ and λ_0 which are not equal to zero simultaneously such that the following conditions (the so-called stationarity conditions)*

$$L'_x(x^*, \lambda^*, \lambda_0^*) = 0 \quad (9.41)$$

hold true. That is,

$$L'_{x_i}(x^*, \lambda^*, \lambda_0^*) = 0, \quad i = 1, 2, \dots, n. \quad (9.42)$$

To guarantee that $\lambda_0 \neq 0$ it is sufficient that the vectors

$$f'_1(x^*), f'_2(x^*), \dots, f'_m(x^*) \quad (9.43)$$

are linearly independent.

Therefore to find the stationary point, one has to solve $n+m$ equations

$$\begin{aligned} \frac{\partial}{\partial x_i} \left(\sum_{k=0}^m \lambda_k f_k(x) \right) &= 0 \quad (\text{n equations, } i = 1, \dots, n) \\ f_1(x) = \dots = f_m(x) &= 0 \quad (m \text{ equations}) \end{aligned} \quad (9.44)$$

with $n+m+1$ unknown values. One must take into account, however, that Lagrange multipliers are defined with accuracy up to a common multiplier. If $\lambda_0 \neq 0$ (this is the most important case since $\lambda_0 = 0$ means that the goal functions are not connected with constraints), then one can multiply all coefficients of the Lagrange multipliers by a constant to obtain $\lambda_0 = 1$. In this case the number of equations becomes equal to the number of unknowns.

One can rewrite Eqs. (9.44) with $\lambda_0 = 1$ in symmetric form:

$$L'_x(x^*, \lambda, 1) = 0,$$

$$L'_{\lambda}(x^*, \lambda, 1) = 0.$$

The solution to these equations defines the stationary points that contain the desired point.

9.5.3 Kuhn-Tucker Theorem (1951)

More than 150 years after Lagrange introduced the multipliers method for solving optimization problems with constraints of equality type, Kuhn and Tucker suggested a solution to the so-called ***convex optimization*** problem, where one minimizes a certain type of (convex) objective function under certain (convex) constraints of inequality type.

Let us remind the reader of the concept of convexity.

Definition. The set A belonging to the linear space is called ***convex*** if along with two points x and y from this set it contains the interval

$$[x, y] = \{z : z = \alpha x + (1 - \alpha)y, \quad 0 \leq \alpha \leq 1\}$$

that connects these points.

Function f is called ***convex*** if for any two points x and y the Jensen inequality

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y), \quad 0 \leq \alpha \leq 1$$

holds true.

Let X be a linear space, let A be a convex subset of this space, and let $f_k(x)$, $k = 0, \dots, m$, be convex functions.

Now consider the following, the so-called ***convex optimization*** problem: Minimize the functional

$$f_0(x) \longrightarrow \inf \tag{9.45}$$

subject to constraints

$$x \in A, \tag{9.46}$$

$$f_k(x) \leq 0, \quad k = 1, \dots, m. \tag{9.47}$$

To solve this problem we consider the Lagrange function (Lagrangian)

$$L = L(x, \lambda_0, \lambda) = \sum_{k=0}^m \lambda_k f_k(x),$$

where $\Lambda = (\lambda_1, \dots, \lambda_m)$. Note that the Lagrangian does not take into account the constraint (9.46).

Theorem (Kuhn-Tucker). *If x^* minimizes function (9.45) under constraints (9.46) and (9.47), then there exist Lagrange multipliers λ_0^* and $\Lambda^* = (\lambda_1^*, \dots, \lambda_m^*)$ that are simultaneously not equal to zero and such that the following three conditions hold true:*

(a) *The minimum principle:*

$$\min_{x \in A} L(x, \lambda_0^*, \mathbf{A}^*) = L(x^*, \lambda_0^*, \mathbf{A}^*). \quad (9.48)$$

(b) *The nonnegativeness conditions:*

$$\lambda_k^* \geq 0, \quad k = 0, 1, \dots, m. \quad (9.49)$$

(c) *The Kuhn–Tucker conditions:*

$$\lambda_k^* f_k(x^*) = 0, \quad k = 1, \dots, m. \quad (9.50)$$

If $\lambda_0 \neq 0$, then conditions (a), (b), and (c) are sufficient conditions for x^* to be the solution of the optimization problem.

In order for $\lambda_0 \neq 0$ it is sufficient that there exists \bar{x} such that the so-called Slater conditions

$$f_i(\bar{x}) < 0, \quad i = 1, \dots, m.$$

hold true.

Corollary. If the Slater condition is satisfied, then one can choose $\lambda_0 = 1$ and rewrite the Lagrangian in the form

$$L(x, 1, \lambda) = f_0(x) + \sum_{k=1}^m \lambda_k f_k(x).$$

Now the Lagrangian defined on $m+n$ variables and conditions of the Kuhn–Tucker theorem are equivalent to the existence of a saddle point (x^*, \mathbf{A}^*) of the Lagrangian, that is,

$$\min_{x \in A} L(x, 1, \lambda^*) = L(x, 1, \mathbf{A}^*) = \max_{\mathbf{A} > 0} L(x^*, 1, \mathbf{A})$$

(minimum taken over $x \in A$ and maximum taken over $\mathbf{A} > 0$)

Indeed the left equality follows from condition (a) of the Kuhn–Tucker theorem, and the right equality follows from conditions (c) and (b).

Note that in the Kuhn–Tucker theorem, condition (a) describes the Lagrange idea: If x^* is the solution of the minimization problem under constraints (9.46) and (9.47), then it provides the minimum of the Lagrange function. Conditions (b) and (c) are specific for constraints of the inequality type.

In the next section we will use Fermat's theorem and the Lagrange multipliers method to derive the so-called back-propagation method for constructing neural nets while in Chapters 10 and 11 we use the Kuhn–Tucker theorem to derive the so-called support vector method for solving a wide range of approximation and estimation problems.

9.6 NEURAL NETWORKS

Up to now, to construct a learning machine we used the following general idea: We (nonlinearly) mapped input vector x in feature space U and then constructed a linear function into this space. In the next chapters we will consider a new idea that will make this approach especially attractive.

However, in the remaining part of this chapter we will come back to the very first idea of the learning machine that was inspired by the neurophysiological analogy. We consider the machine defined by a superposition of several neurons. This structure has n input and one output and is defined by connections of several neurons each with their own weights. This construction is called a Multilayer Perceptron or Neural Network. Learning in neural networks is the same as estimating the coefficients of all neurons. To estimate these coefficients, one considers the model of neurons where instead of the threshold function one uses a sigmoid function.

As we demonstrate in Section 9.3 to define the procedure for estimating the unknown coefficients (weights) for all neurons, it is sufficient to calculate the gradient of the loss function for the neural networks.

The method for calculating the gradient of loss function for the sigmoid approximation of neural networks, called the back-propagation *method*, was proposed in 1986 (Rumelhart, Hinton, and Williams, 1986; LeCun, 1986).

Using gradients, one can iteratively modify the coefficients (weights) of a neural net on the basis of standard gradient-based procedures.

9.6.1 The Back-Propagation Method

To describe the back-propagation method we use the following notations (Fig. 9.2):

1. The neural net contains $m + 1$ layers connected each to other: The first layer $x(0)$ describes the input vector $x = (x^1, \dots, x^n)$. We denote the input vector by

$$x_i = (x_i^1(0), \dots, x_i^n(0)), \quad i = 1, \dots, \ell,$$

and we denote the image of the input vector $x_i(0)$ on the k th layer by

$$x_i(k) = (x_i^1(k), \dots, x_i^{n_k}(k)), \quad i = 1, \dots, \ell,$$

where we denote by n_k the dimensionality of the vectors $x_i(k)$, $i = 1, \dots, \ell$ (n_k , $k = 1, \dots, m - 1$, can be any number, but $n_m = 1$).

2. The layer $k - 1$ is connected with the layer k through the $(n_{k-1} \times n_k)$ matrix $w(k)$ as follows:

$$x_i(k) = S\{w(k)x_i(k - 1)\}, \quad k = 1, 2, \dots, m, \quad i = 1, \dots, \ell. \quad (9.51)$$

In Eq. (9.51) we use the following notations: Vector $S\{\mathbf{w}(k)x_i(k-1)\}$ is defined by the sigmoid $S(\mathbf{u})$ and the vector

$$\mathbf{u}_i(k) = (u_i^1(k), \dots, u_i^{n_k}(k)) = \mathbf{w}(k)x_i(k-1),$$

where the sigmoid function transforms the coordinates of the vector:

$$S(\mathbf{u}_i(k)) = (S(u_i^1(k)), \dots, S(u_i^{n_k}(k))).$$

The goal is to minimize the functional

$$R(\mathbf{w}(1), \dots, \mathbf{w}(m)) = \sum_{i=1}^{\ell} (y_i - x_i(m))^2 \quad (9.52)$$

under conditions (9.51).

This optimization problem is solved by using the standard technique of Lagrange multipliers for equality type constraints. We will minimize the Lagrange function

$$\begin{aligned} L(W, X, B) \\ = \sum_{i=1}^{\ell} (y_i - x_i(m))^2 + \sum_{i=1}^{\ell} \sum_{k=1}^m (b_i(k) * [x_i(k) - S\{\mathbf{w}(k)x_i(k-1)\}]), \end{aligned}$$

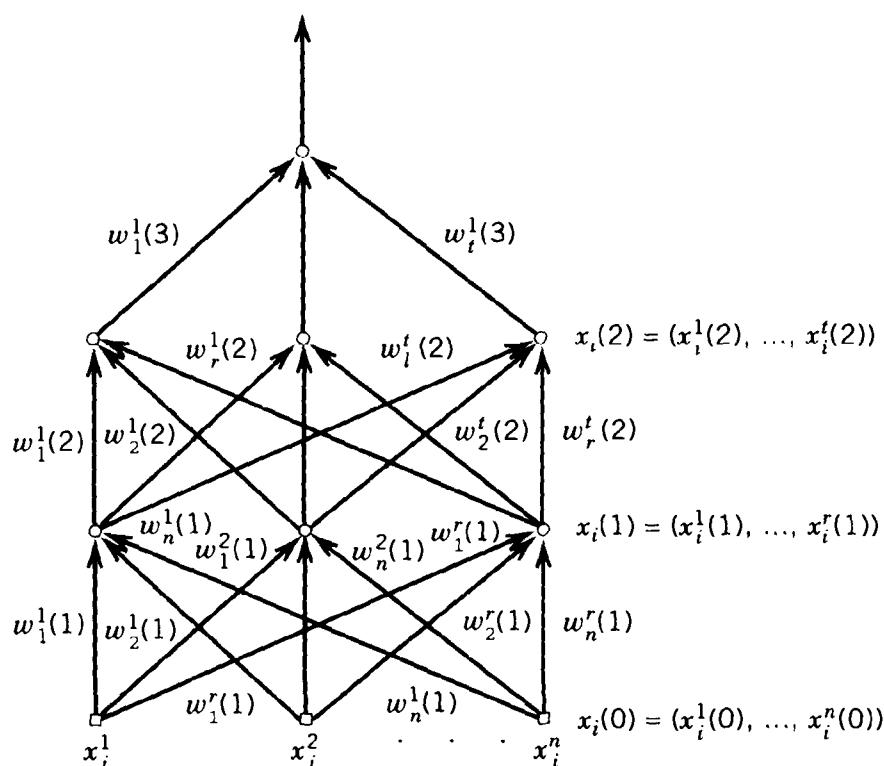


FIGURE 9.2. A neural network is a combination of several levels of sigmoid elements. The outputs of one layer form the input for the next layer.

where $b_i(k)$ are Lagrange multipliers corresponding to the constraints (9.51) that describe the connections between vectors $x_i(k - 1)$ and vectors $x_i(k)$.

The equality

$$\nabla L(W, X, B) = 0$$

is a necessary condition for a local minimum of the performance function (9.52) under the constraints (9.51) (the gradient with respect to all parameters from $b_i(k)$, $x_i(k)$, $w(k)$, $i = 1, \dots, \ell$, $k = 1, \dots, m$, is equal to zero).

This condition can be split into three subconditions:

1. $\frac{\partial L(W, X, B)}{\partial b_i(k)} = 0, \quad \forall i, k.$
2. $\frac{\partial L(W, X, B)}{\partial x_i(k)} = 0, \quad \forall i, k.$
3. $\frac{\partial L(W, X, B)}{\partial w(k)} = 0 \quad \forall w(k).$

The solution of these equations determines a stationary point (W_0, X_0, B_0) that includes the desired matrices of weights $W_0 = (w^0(1), \dots, w^0(m))$. Let us rewrite these three subconditions in explicit form:

1. The First Subcondition. The first subcondition gives a set of equations

$$x_i(k) = S \{w(k)x_i(k - 1)\}, \quad i = 1, \dots, \ell, \quad k = 1, \dots, m$$

with initial conditions

$$x_i(0) = x_i,$$

the equation of the so-called forward dynamics. It iteratively defines images of the input vectors $x_i(0)$ for all levels $k = 1, \dots, m$ of the neural net.

2. The Second Subcondition. We consider the second subconditions for two cases: for the case $k = m$ (for the last layer) and for the case $k \neq m$ (for hidden layers).

For the last layer we obtain

$$b_i(m) = 2(y_i - x_i(m)), \quad i = 1, \dots, \ell.$$

For the general case (hidden layers) we obtain

$$b_i(k) = w^T(k + 1) \nabla S \{w(k + 1)x_i(k)\} b_i(k + 1), \\ i = 1, \dots, \ell, \quad k = 1, \dots, m - 1,$$

where $\nabla S \{w(k + 1)x_i(k)\}$ is a diagonal $n_{k+1} \times n_{k+1}$ matrix with diagonal elements $S'(u^r)$, where u^r is the r th coordinate of the (n_{k+1} -dimensional) vector $u = w(k + 1)x_i(k)$. This equation describes the backward dynamics. It iteratively defines Lagrange multipliers $b_i(k)$ for all $k = m, \dots, 1$ and all $i = 1, \dots, C$.

3. *The Third Subcondition.* Unfortunately the third subcondition does not give a direct method for computing the matrices of weights $w(k)$, $k = 1, \dots, m$. Therefore to estimate the weights, one uses steepest gradient descent:

$$w(k) \leftarrow w(k) - \gamma_t \frac{\partial L(W, X, B)}{\partial w(k)}, \quad k = 1, \dots, m,$$

where γ_t is a value of step for iteration t . In explicit form, this equation is

$$w(k) \leftarrow w(k) + \gamma_t \sum_{i=1}^{\ell} \nabla S \{w(k)x_i(k-1)\} b_i(k)x_i^T(k-1), \\ k = 1, 2, \dots, m.$$

This equation describes the rule for iterative weight updating.

9.6.2 The Back-Propagation Algorithm

Therefore the back-propagation algorithm contains three elements:

1. *Forward Pass:*

$$x_i(k) = S \{w(k)x_i(k-1)\}, \quad i = 1, \dots, \ell, \quad k = 1, \dots, m$$

with the boundary conditions

$$x_i(0) = x_i, \quad i = 1, \dots, \ell.$$

2. *Backward Pass:*

$$b_i(k) = w^T(k+1) \nabla S \{w(k+1)x_i(k)\} b_i(k+1),$$

$$i = 1, \dots, \ell, \quad k = 1, \dots, m-1$$

with the boundary conditions

$$b_i(m) = 2(y_i - x_i(m)), \quad i = 1, \dots, \ell.$$

3. *Weight Update for Weight Matrices $w(k)$, $k = 1, 2, \dots, m$:*

$$w(k) \leftarrow w(k) + \gamma_t \sum_{i=1}^{\ell} \nabla S \{w(k)x_i(k-1)\} b_i(k)x_i^T(k-1),$$

Using the back-propagation technique one can achieve a local minimum for the empirical risk functional.

9.6.3 Neural Networks for the Regression Estimation Problem

To adapt neural networks for solving the regression estimation problem, it is sufficient to use in the last layer a linear function instead of a sigmoid one. This implies only the following changes in the above equations:

$$x_i(m) = w(m)x_i(m-1), \quad i = 1, \dots, \ell,$$

$$\nabla S\{w(m)x_i(m-1)\} = 1.$$

9.6.4 Remarks on the Back-Propagation Method

The main problems with the neural net approach are as follows:

1. The empirical risk functional has many local minima. Standard optimization procedures guarantee convergence to one of them. The quality of the obtained solution depends on many factors, in particular on the initialization of weight matrices $w(k)$, $k = 1, \dots, m$. The choice of initialization parameters to achieve a "small" local minimum is based on heuristics.
2. The convergence of the gradient based method is rather slow. There are several heuristics to speed up the rate of convergence.
3. The sigmoid function has a scaling factor that affects the quality of the approximation. The choice of the scaling factor is a trade-off between the quality of approximation and the rate of convergence. There are empirical recommendations for choosing the scaling factor.

Therefore neural networks are not well-controlled learning machines. Nevertheless, in many practical applications, neural networks demonstrate good results.

THE SUPPORT VECTOR METHOD FOR ESTIMATING INDICATOR FUNCTIONS

Chapter 9 showed that methods of separating hyperplanes play an important role in constructing learning algorithms. These methods were the foundation of classical learning algorithms.

This chapter considers a special type of separating hyperplanes, the so-called optimal hyperplanes, that possess some remarkable statistical properties. Using the method of the optimal separating hyperplane we construct a new class of learning machines for estimating indicator functions, the so-called support vector machines, which we will generalize in the next chapter for estimating real-valued functions, signal processing, and solving linear operator equations.

10.1 THE OPTIMAL HYPERPLANE

We say that two finite subsets of vectors \mathbf{x} from the training set

$$(y_1, \mathbf{x}_1), \dots, (y_\ell, \mathbf{x}_\ell), \quad \mathbf{x} \in R^n, \quad y \in \{-1, 1\},$$

one subset I for which $y = 1$, and another subset II for which $y = -1$ are separable by the hyperplane

$$(\mathbf{x} * \phi) = c$$

if there exist both a unit vector ϕ ($|\phi| = 1$) and a constant c such that the inequalities

$$\begin{aligned} (\mathbf{x}_i * \phi) &> c, & \text{if } \mathbf{x}_i \in I, \\ (\mathbf{x}_j * \phi) &< c, & \text{if } \mathbf{x}_j \in II \end{aligned} \tag{10.1}$$

hold true where we denoted by $(a * b)$ the inner product between vectors a and b .

Let us determine for any unit vector ϕ two values

$$c_1(\phi) = \min_{x_i \in I} (x_i * \phi),$$

$$c_2(\phi) = \max_{x_j \in II} (x_j * \phi).$$

Consider the unit vector ϕ_0 which maximizes the function

$$\rho(\phi) = \frac{c_1(\phi) - c_2(\phi)}{2}, \quad |\phi| = 1 \quad (10.2)$$

under the condition that inequalities (10.1) are satisfied. The vector ϕ_0 and the constant

$$c_0 = \frac{c_1(\phi_0) + c_2(\phi_0)}{2}$$

determine the hyperplane that separates vectors x_1, \dots, x_a of the subset I from vectors x_1, \dots, x_b of the subset II, ($a+b=\ell$) and has the maximal margin (10.2). We call this hyperplane the ***maximal margin hyperplane*** or the ***optimal hyperplane*** (Fig. 10.1).

Theorem 10.1. *The optimal hyperplane is unique.*

Proof. We need to show that the maximum point ϕ_0 of the continuous function $\rho(\phi)$ defined in the area $|\phi| \leq 1$ exists and is achieved on the boundary $|\phi| = 1$. Existence of the maximum follows from the continuity of $\rho(\phi)$ in the bounded region $|\phi| \leq 1$.

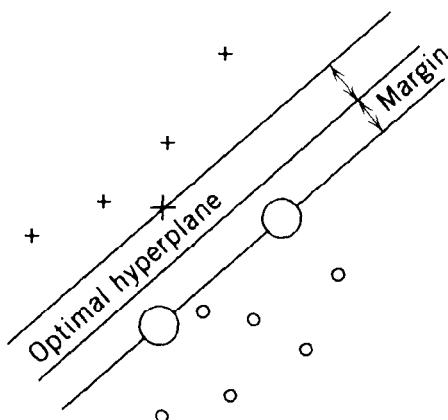


FIGURE 10.1. The optimal separating hyperplane is the one that separates the data with the maximal margin.

Suppose that the maximum is achieved at some interior point ϕ^* . Then the vector

$$\phi^* = \frac{\phi_0}{|\phi_0|}$$

would define a larger margin

$$\rho(\phi^*) = \frac{\rho(\phi_0)}{|\phi_0|}.$$

The maximum of the function $\rho(\phi)$ cannot be achieved on two (boundary) points. Otherwise, since function $\rho(\phi)$ is convex, it is achieved on the line that connects these two points—that is, at an inner point which is impossible by preceding arguments. This proves the theorem.

Our goal is to find effective methods for constructing the optimal hyperplane. To do so we consider an equivalent statement of the problem: Find a pair consisting of a vector ψ_0 and a constant (threshold) b_0 such that they satisfy the constraints

$$\begin{aligned} (x_i * \psi_0) + b_0 &\geq 1, & \text{if } y_i = 1, \\ (x_j * \psi_0) + b_0 &\leq -1, & \text{if } y_j = -1, \end{aligned} \quad (10.3)$$

and the vector ψ_0 has the smallest norm

$$|\psi|^2 = (\psi * \psi). \quad (10.4)$$

Theorem 10.2. *Vector ψ_0 that minimizes (10.4) under constraints (10.3) is related to the vector that forms the optimal hyperplane by the equality*

$$\phi_0 = \frac{\psi_0}{|\psi_0|}. \quad (10.5)$$

The margin ρ_0 between the optimal hyperplane and separated vectors is equal to

$$\rho(\phi_0) = \sup_{\phi_0} \frac{1}{2} \left(\min_{i \in I} (x_i * \phi_0) - \max_{j \in II} (x_j * \phi_0) \right) = \frac{1}{|\psi_0|}. \quad (10.6)$$

Proof. Indeed, the vector ψ_0 that provides the minimum of the quadratic function (10.4) under the linear constraints (10.3) is unique. Let us define the unit vector

$$\phi_0 = \frac{\psi_0}{|\psi_0|}.$$

Since constraints (10.3) are valid, we find

$$\rho \left(\frac{\psi_0}{|\psi_0|} \right) = \frac{1}{2} \left(c_1 \left(\frac{\psi_0}{|\psi_0|} \right) - c_2 \left(\frac{\psi_0}{|\psi_0|} \right) \right) \geq \frac{1}{|\psi_0|}.$$

To prove the theorem it is sufficient to show that the inequality

$$\rho \left(\frac{\psi_0}{|\psi_0|} \right) > \frac{1}{|\psi_0|}$$

is impossible. Suppose it holds true. Then there exists a unit vector ϕ^* such that the inequality

$$\rho(\phi^*) > \frac{1}{|\psi_0|}$$

holds true. Let us construct the new vector

$$\psi^* = \frac{\phi^*}{\rho(\phi^*)},$$

which has norm smaller than $|\psi_0|$. One can check that this vector satisfies the constraints (10.3)with

$$b = -c_1(\phi) + c_2(\phi)$$

This contradicts the assertion that ψ_0 is the smallest vector satisfying the constraints (10.3).This proves the theorem.

Thus the vector ψ_0 with the smallest norm satisfying constraints (10.3)defines the optimal hyperplane. The vector ψ_0 with the smallest norm satisfying constraints (10.3)with $b = 0$ defines the optimal hyperplane passing through the origin.

To simplify our notation let us rewrite the constraint (10.3)in the equivalent form

$$y_i((x_i * \psi_0) + b) \geq 1, \quad i = 1, \dots, \ell. \quad (10.7)$$

Therefore in order to find the optimal hyperplane one has to solve the following quadratic optimization problem: to minimize the quadratic form (10.4) subject to the linear constraints (10.7).

One can solve this problem in the primal space—the space of parameters ψ and \mathbf{b} . However, the deeper results can be obtained by solving this quadratic optimization problem in the dual space—the space of Lagrange multpliers. Below we consider this type of solution.

As it was shown in Section 9.5, in order to solve this quadratic optimization problem one has to find the saddle point of the Lagrange function

$$L(\psi, b, \alpha) = \frac{1}{2}(\psi * \psi) - \sum_{i=1}^{\ell} \alpha_i (y_i[(x_i * \psi) + b] - 1), \quad (10.8)$$

where $\alpha_i \geq 0$ are the Lagrange multipliers. To find the saddle point one has to minimize this function over ψ and b and to maximize it over the nonnegative Lagrange multipliers $a, \geq 0$.

According to the Fermat theorem, the minimum points of this functional have to satisfy the conditions

$$\begin{aligned}\frac{\partial L(\psi, b, \alpha)}{\partial \psi} &= \psi - \sum_{i=1}^{\ell} y_i \alpha_i x_i = 0, \\ \frac{\partial L(\psi, b, \alpha)}{\partial b} &= \sum_{i=1}^{\ell} y_i \alpha_i = 0.\end{aligned}$$

From these conditions it follows that for the vector ψ that defines the optimal hyperplane, the equalities

$$\psi = \sum_{i=1}^{\ell} y_i \alpha_i x_i, \quad (10.9)$$

$$\sum_{i=1}^{\ell} y_i \alpha_i = 0 \quad (10.10)$$

hold true. Substituting (10.9) into (10.8) and taking into account (10.10), one obtains

$$W(\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j (x_i * x_j). \quad (10.11)$$

Note that we have changed the notation from $L(\psi, b, \alpha)$ to $W(\alpha)$ to reflect the last transformation. Now to construct the optimal hyperplane one has to find the coefficients α_i^0 that maximize the function (10.11) in the nonnegative quadrant

$$\alpha_i \geq 0, \quad i = 1, \dots, \ell, \quad (10.12)$$

under the constraint (10.10). Using these coefficients α_i^0 , $i = 1, \dots, \ell$, in Eq. (10.9), one obtains the solution

$$\psi_0 = \sum_{i=1}^{\ell} y_i \alpha_i^0 x_i.$$

The value of b_0 is chosen to maximize margin 10.2. Note that the optimal solution ψ_0 and b_0 must satisfy the Kuhn–Tucker conditions (see Chapter 9, Section 9.5)

$$\alpha_i^0 (y_i ((x_i * \psi_0) + b_0) - 1) = 0, \quad i = 1, \dots, \ell. \quad (10.13)$$

From conditions (10.13) one concludes that nonzero values α_i^0 correspond only to the vectors x_i that satisfy the equality

$$y_i((x_i * \psi_0) + b_0) = 1. \quad (10.14)$$

Geometrically, these vectors are the closest to the optimal hyperplane (see Fig 10.1). We will call them *support vectors*. The support vectors play a crucial role in constructing a new type of learning algorithm since the vector ψ_0 that defines the optimal hyperplane is expanded with nonzero weights on support vectors:

$$\psi_0 = \sum_{i=1}^{\ell} y_i \alpha_i^0 x_i.$$

Therefore the optimal hyperplane has the form

$$f(x, \alpha_0) = \sum_{i=1}^{\ell} y_i \alpha_i^0 (x_s * x) + b_0, \quad (10.15)$$

where $(x, * x)$ is the inner product of the two vectors.

Note that both the separation hyperplane (10.15) and the objective function of our optimization problem

$$W(\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j (x_i * x_j) \quad (10.16)$$

do not depend explicitly on the dimensionality of the vector x but depend only on the inner product of two vectors. This fact will allow us later to construct separating hyperplanes in high-dimensional spaces even (in infinite-dimensional Hilbert spaces).

We now formulate some properties of the optimal hyperplane that are used later.

1. The optimal hyperplane is unique; that is, the pair of vector ψ_0 and threshold b_0 that define the optimal hyperplane is unique. However, the expansion of the vector ψ_0 on the support vectors is not unique.
2. Let the vector ψ_0 define the optimal hyperplane. Then the maximum of the functional $W(\alpha)$ is equal to

$$W(\alpha_0) = \frac{1}{2} (\psi_0 * \psi_0) = \frac{1}{2} \sum_i \alpha_i^0.$$

To show this, it is sufficient to transform functional (10.11), taking into account that for optimal pairs ψ_0 and b_0 the equality (10.10) and equalities (10.13) hold true. This implies

$$(\psi_0 * \psi_0) = \sum_i \alpha_i^0$$

3. The norm of the vector ψ_0 defines the margin of the optimal separating hyperplane

$$\rho(\psi_0) = \frac{1}{|\psi_0|}.$$

4. From properties 2 and 3 it follows that

$$W(\alpha) < W(\alpha_0) = \frac{1}{2} \left(\frac{1}{\rho(\psi_0)} \right)^2, \quad \alpha \neq \alpha_0.$$

This expression can be chosen as a criterion of linear nonseparability of two sets of data.

Definition. We call two sets of data linearly 8-nonseparable if the margin between the hyperplane and the closest vector is less than 8.

Therefore, if during the maximization procedure the value $W(\alpha)$ exceeds the value $1/2S^2$, one can assert that the two sets of separating data are 8-nonseparable.

Thus, in order to construct the optimal hyperplane, one has either to find the maximum of the nonnegative quadratic form $W(\alpha)$ in the nonnegative quadrant under the constraint (10.10) or to establish that the maximum exceeds the value

$$W_{\max} = \frac{1}{2\delta^2}.$$

In the latter case, separation with the margin S is impossible.

5. In order to maximize the functional $W(\alpha)$ under the constraints (10.10) and (10.12), one needs to specify the support vectors and to determine the corresponding coefficients. This can be done sequentially, using a small amount of training data every time. One can start the optimization process using only n examples (maximizing $W(\alpha)$ under the condition that only n parameters differ from zero). As the conditional maximum $W(\alpha)$ is achieved, one can keep the parameters that differ from zero and add new parameters (corresponding to the vectors that were not separated correctly by the first iteration of constructing the optimal hyperplane). One continues this process until either:

- (a) all the vectors of the training set are separated, or
- (b) at some step $W(\alpha) > W_{\max}$ (the separation is impossible).

The methods described above work in some sense like a sieve: At any step it maximizes the functional $W(\alpha)$ using only the elements of the training set which are the candidates for support vectors.

10.2 THE OPTIMAL HYPERPLANE FOR NONSEPARABLE SETS

10.2.1 The Hard Margin Generalization of the Optimal Hyperplane

In this section we generalize the concept of the optimal hyperplane for the nonseparable case.

Let the set of training set

$$(y_1, x_1), \dots, (y_\ell, x_\ell), \quad x \in X, y \in \{-1, 1\},$$

be such that it cannot be separated without error by a hyperplane. According to our definition of nonseparability (see Section 10.1), this means that there is no pair ψ, b such that

$$(\psi * \psi) \leq \frac{1}{\rho^2} = A^2$$

and the inequalities

$$y_i((x_i * \psi) + b) \geq 1, \quad i = 1, 2, \dots, \ell \quad (10.17)$$

hold true.

Our goal is to construct the hyperplane that makes the smallest number of errors. To get a formal setting of this problem we introduce the nonnegative variables

$$\xi_1, \dots, \xi_\ell.$$

In terms of these variables the problem of finding the hyperplane that provides the minimal number of training errors has the following formal expression: Minimize the functional

$$\Phi(\xi) = \sum_{i=1}^{\ell} \theta(\xi_i)$$

subject to the constraints

$$y_i((x_i * \psi) + b) \geq 1 - \xi_i, \quad i = 1, 2, \dots, \ell, \quad \xi_i \geq 0 \quad (10.18)$$

and the constraint

$$(\psi * \psi) \leq A^2, \quad (10.19)$$

where $\delta(\xi) = 0$ if $\xi = 0$ and $\theta(\xi) = 1$ if $\xi > 0$. It is known that for the nonseparable case this optimization problem is NP-complete. Therefore we consider the following approximation to this problem: We would like to minimize the functional

$$\Phi(\xi) = \sum_{i=1}^{\ell} \xi_i^\sigma$$

under the constraints (10.18) and (10.19), where $\sigma \geq 0$ is a small value. We will, however, choose $\sigma = 1$, the smallest σ that leads to a simple optimization problem.[†]

Thus, we will minimize the functional

$$\Phi(\psi, b) = \sum_{i=1}^{\ell} \xi_i \quad (10.20)$$

subject to the constraints (10.18) and (10.19). We call the hyperplane

$$(\psi_0 * x) + b = 0$$

constructed on the basis of the solution of this optimization problem the *generalized optimal hyperplane* or, for simplicity, the *optimal hyperplane*.

To solve this optimization problem we find the saddle point of the Lagrangian

$$\begin{aligned} L(\psi, b, \alpha, \beta, \gamma) = & \sum_{i=1}^{\ell} \xi_i - \frac{1}{2} \gamma (A^2 - (\psi * \psi)) \\ & - \sum_{i=1}^{\ell} \alpha_i (y_i ((\psi * x_i) + b) - 1 + \xi_i) - \sum_{i=1}^{\ell} \beta_i \xi_i \end{aligned} \quad (10.21)$$

(the minimum with respect to ψ , b , ξ_i and the maximum with respect to non-negative multipliers α_i , β_i , y). The parameters that minimize the Lagrangian

[†]The choice $a = 2$ also leads to a simple optimization problem. However, for the pattern recognition problem this choice does not look attractive. It will be more attractive when we will generalize results obtained for the pattern recognition problem to estimation of real-valued functions (Chapter 11, Section 11.3).

must satisfy the conditions

$$\begin{aligned}\frac{\partial L(\psi, b, \xi, \alpha, \beta, \gamma)}{\partial \psi} &= \gamma \psi - \sum_{i=1}^{\ell} \alpha_i y_i x_i = 0, \\ \frac{\partial L(\psi, b, \xi, \alpha, \beta, \gamma)}{\partial b} &= - \sum_{i=1}^{\ell} y_i \alpha_i = 0, \\ \frac{\partial L(\psi, b, \xi, \alpha, \beta, \gamma)}{\partial \xi_i} &= 1 - \alpha_i - \beta_i = 0.\end{aligned}$$

From these conditions one derives

$$\psi = \frac{1}{\gamma} \sum_{i=1}^{\ell} \alpha_i y_i x_i, \quad (10.22)$$

$$\begin{aligned}\sum_{i=1}^{\ell} \alpha_i y_i &= 0, \\ \alpha_i + \beta_i &= 1.\end{aligned} \quad (10.23)$$

Substituting (10.22) into the Lagrangian and taking into account (10.23), we obtain the functional

$$W(\alpha, \gamma) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2\gamma} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j (x_i * x_j) - \frac{\gamma A^2}{2}, \quad (10.24)$$

which one has to maximize under the constraints

$$\begin{aligned}\sum_{i=1}^{\ell} y_i \alpha_i &= 0, \\ 0 \leq \alpha_i &\leq 1, \\ \gamma &\geq 0.\end{aligned}$$

One can maximize (10.24) under these constraints by solving a quadratic optimization problem several times for fixed values of y and conducting maximization with respect to y by a line search. One can also find the parameter y that maximizes (10.24) and substitute it back into (10.24). It is easy to check that the maximum of (10.24) is achieved when

$$\gamma = \frac{\sqrt{\sum_{i,j}^{\ell} \alpha_i \alpha_j y_i y_j (x_i * x_j)}}{A}$$

Putting this expression back into (10.24), one obtains that to find the desired hyperplane one has to maximize the functional

$$W(\alpha) = \sum_{i=1}^{\ell} \alpha_i - A \sqrt{\sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j (x_i * x_j)} \quad (10.25)$$

subject to constraints

$$\begin{aligned} \sum_{i=1}^{\ell} y_i \alpha_i &= 0, \\ 0 \leq \alpha_i &\leq 1. \end{aligned} \quad (10.26)$$

The vector of parameters $\alpha_0 = (\alpha_1^0, \dots, \alpha_\ell^0)$ defines the generalized optimal hyperplane

$$f(x) = \frac{A}{\sqrt{\sum_{i,j=1}^{\ell} \alpha_i^0 \alpha_j^0 y_i y_j (x_i * x_j)}} \sum_{i=1}^{\ell} \alpha_i^0 y_i (x * x_i) + b.$$

The value of the threshold b is chosen to satisfy the Kuhn–Tucker condition

$$\alpha_t^0 \left(\frac{A}{\sqrt{\sum_{i,j=1}^{\ell} \alpha_i^0 \alpha_j^0 y_i y_j (x_i * x_j)}} \sum_{i=1}^{\ell} \alpha_i^0 y_i (x_t * x_i) + b \right) = 0, \quad t = 1, \dots, \ell.$$

10.2.2 The Basic Solution. Soft Margin Generalization

To simplify computations, one can introduce the following (slightly modified) concept of the generalized optimal hyperplane. The generalized optimal hyperplane is determined by the vector ψ that minimizes the functional

$$\Phi(\psi, \xi) = \frac{1}{2}(\psi * \psi) + C \left(\sum_{i=1}^{\ell} \xi_i \right) \quad (10.27)$$

subject to the constraints (10.17) (here C is a given value).

Using the same technique with the Lagrangian, one obtains the method for solution of this optimization problem that is almost equivalent to the method of solution of the optimization problem for the separable case: To find the vector ψ of the generalized optimal hyperplane

$$\psi = \sum_{i=1}^{\ell} \alpha_i y_i x_i,$$

one has to maximize the same quadratic form as in the separable case

$$W(\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j (x_i * x_j) \quad (10.28)$$

under slightly different constraints:

$$\begin{aligned} 0 \leq \alpha_i \leq C, \quad i = 1, \dots, \ell, \\ \sum_{i=1}^{\ell} \alpha_i y_i = 0. \end{aligned} \quad (10.29)$$

As in the separable case, only some of the coefficients α_i^0 , $i = 1, \dots, \ell$, differ from zero. They and the corresponding support vectors determine the generalized optimal separating hyperplane

$$\sum_{i=1}^{\ell} \alpha_i^0 y_i (x_i * x) + b_0 = 0. \quad (10.30)$$

Note that if the coefficient C in the functional (10.27) is equal to the optimal value of parameter γ_0 for maximization of the functional (10.24)

$$C = Y_0,$$

then the solutions of both optimization problems coincide.

10.3 STATISTICAL PROPERTIES OF THE OPTIMAL HYPERPLANE

This section discusses some statistical properties of the optimal hyperplane. In particular, we discuss theorems showing that the bounds on generalization ability of the optimal hyperplane are better than the general bounds obtained for method minimizing the empirical risk.

Let $X^* = (x_1, \dots, x_\ell)$ be a set of ℓ vectors in \mathbf{R}^n . For any hyperplane

$$(x * \psi^*) + b^* = 0$$

in \mathbf{R}^n consider the corresponding *canonical hyperplane* defined by the set X^* as follows:

$$\inf_{x \in X^*} |(x * \psi) + b| = 1,$$

where $\psi = c^* \psi^*$ and $b = c^* b^*$. Note that the set of canonical hyperplanes coincides with the set of separating hyperplanes. It only specifies the normalization with respect to given set of data X^* .

First let us establish the following important fact.

Theorem 10.3. A subset of canonical hyperplane defined on $X^* \subset R^n$

$$|x| \leq D, \quad x \in X^*$$

satisfying the constraint

$$|\psi| \leq A$$

has the VC dimension h bounded as follows:

$$h \leq \min ([D^2 A^2], n) + 1,$$

where $[a]$ denotes the integer part of a .

Note that the norm of the vector coefficients of the canonical hyperplane ψ defines the margin

$$\rho\left(\frac{\psi}{|\psi|}\right) = \frac{c_1\left(\frac{\psi}{|\psi|}\right) - c_2\left(\frac{\psi}{|\psi|}\right)}{2} = \frac{1}{|\psi|} = \frac{1}{A}$$

(see Section 10.1).

Therefore when $[D^2 A^2] < n$ the VC dimension is bounded by the same construction D^2/ρ^2 that in Theorem 9.1 defines the number of corrections made by the Perceptron. This time the construction D^2/ρ^2 is used to bound the VC dimension of the set of hyperplanes with margin not less than p defined on the set of vectors X^* .

To formulate the next theorems we introduce one more concept. The last section mentioned that the minimal norm vector ψ satisfying the conditions (10.7) is unique, though it can have different expansions on the support vectors.

Definition 2. We define the support vectors x_i that appear in all possible expansions of the vector ψ_0 the *essential support vectors*. In other words, the essential support vectors comprise the intersection of all possible sets of support vectors. Let

$$(x_1, y_1), \dots, (x_\ell, y_\ell)$$

be the training set. We denote the number of essential support vectors of this training set by

$$\mathcal{K}_\ell = \mathcal{K}((x_1, y_1), \dots, (x_\ell, y_\ell)).$$

We also denote the maximum norm of the vector x from a set of essential support vectors of this training set by

$$D_\ell = D((x_1, y_1), \dots, (x_\ell, y_\ell)) = \max_i |x_i|.$$

Let n be the dimensionality of the vectors x .

The following four theorems describe the main statistical properties of the optimal hyperplanes.

Theorem 10.4. *The inequality*

$$\mathcal{K}_\ell \leq n \quad (10.31)$$

holds true.

Theorem 10.5. *Let*

$$ER(\alpha_\ell) = ER(y, x, \alpha(y_1, x_1, \dots, y_\ell, x_\ell))$$

be the expectation of the probability of error for optimal hyperplanes constructed on the basis of training samples of size ℓ (the expectation taken over both training and test data). Then the following inequality

$$ER(\alpha_\ell) \leq \frac{E\mathcal{K}_{\ell+1}}{\ell + 1} \quad (10.32)$$

holds true.

Corollary. *From Theorems 10.4 and 10.5 it follows that*

$$ER(\alpha_\ell) \leq \frac{n}{\ell + 1}.$$

Theorem 10.6. *For optimal hyperplanes passing through the origin the following inequality*

$$ER(\alpha_\ell) \leq \frac{E \left(\frac{\mathcal{D}_{\ell+1}}{\rho_{\ell+1}} \right)^2}{\ell + 1} \quad (10.33)$$

holds true, where $\mathcal{D}_{\ell+1}$ and $\rho_{\ell+1}$ are (random) values that for a given training set of size $\ell + 1$ define the maximal norm of support vectors x and the margin.

Remark. In Section 9.1, while analyzing the Perceptron's algorithm, we discussed the Novikoff theorem which bounds the number M of error corrections that the Perceptron makes in order to construct a separating hyperplane. The bound is

$$M \leq \left[\frac{D_\ell^2}{\rho_\ell^2} \right]$$

where D_ℓ is the bound on the norm of vectors on which the correction was made and ρ_ℓ is the maximal possible margin between the separating hyperplane and the closest vector to the hyperplane.

In Section 10.4.4 along with Theorem 10.5, we prove Theorem 9.3 which states that after separating training data (may be using them several times) the Perceptron constructs a separate hyperplane whose error has the following bound

$$ER(w_\ell) \leq \frac{E \min \left\{ \left[\frac{D_{\ell+1}^2}{\rho_{\ell+1}^2} \right], K \right\}}{\ell + 1} \quad (10.34)$$

where K is the number of correction made by the Perceptron. Compare this bound with the bound obtained in Theorem 10.6 for the optimal separating hyperplane. The bounds have the same structure and the same value ρ_ℓ^2 of the margin. The only difference is that the D_ℓ^2 (the bound on the norm for support vectors) in Theorem 10.6 for optimal separating hyperplanes is used instead of D_ℓ (the bound on the norm for correcting vectors) in (10.34) for the Perceptron's hyperplanes.

In these bounds the advantage of comparing optimal hyperplanes to Perceptron's hyperplanes is expressed through the geometry of support vectors.

Theorem 10.7. For the optimal hyperplane passing through the origin *the* inequality

$$ER(\alpha_\ell) \leq \frac{E \min \left(K_{\ell+1}, \left(\frac{D_{\ell+1}}{\rho_{\ell+1}} \right)^2 \right)}{\ell + 1}$$

is valid.

10.4 PROOFS OF THE THEOREMS

10.4.1 Proof of Theorem 10.3

To estimate the VC dimension of the canonical hyperplane, one has to estimate the maximal number r of vectors that can be separated in all 2^r possible ways by hyperplanes with the margin p . This bound was obtained in Theorem 8.4. Therefore the proof of this theorem coincides with the proof of Theorem 8.4 given in Chapter 8, Section 8.5.

10.4.2 Proof of Theorem 10.4

Let us show that the number of essential support vectors does not exceed the dimensionality of the space X . To prove this we show that if an expansion of the optimal hyperplane has a $> n$ support vectors (with nonzero coefficients), then there exists another expansion that contains less support vectors.

Indeed, suppose that the optimal hyperplane is expanded as follows:

$$\psi = \sum_{i=1}^a \alpha_i x_i y_i, \quad (10.35)$$

where $a > n$. Since any system of vectors that contains $a > n$ different elements is linearly dependent, there exist parameters γ_i such that

$$\sum_{i=1}^a \gamma_i x_i y_i = 0,$$

where at least one γ_i is positive.

Therefore the expression

$$\psi = \sum_{i=1}^{\alpha} (\alpha_i - t \gamma_i) x_i y_i \quad (10.36)$$

determines a family of expansions of the optimal hyperplanes depending on t . Since all α_i are positive, all the coefficients remain positive for sufficiently small t . On the other hand, since among the coefficients γ_i some are positive, for sufficiently large $t > 0$, some coefficients become negative. Therefore one can find such minimal $t = t_0$ for which one or several coefficients become zero for the first time while the rest of the coefficients remain positive. Taking the value of $t = t_0$, one can find an expansion with a reduced number of support vectors.

Utilizing this construction several times, one can obtain an expansion of the optimal hyperplane which is based on at most n vectors.

10.4.3 Leave-One-Out Procedure

The proofs of the next two theorems are based on the so-called leave-one-out estimation procedure. Below we use this procedure as a tool for proving our theorems, although this procedure is usually used for evaluating the probability of test error of the function obtained by the empirical risk minimization method.

Let $Q(z, \alpha_\ell)$ be the function that minimizes the empirical risk

$$R_{\text{emp}}(\alpha) = \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \quad (10.37)$$

on a given sequence

$$z_1, \dots, z_\ell. \quad (10.38)$$

Let us estimate the risk for the function $Q(z, \alpha_\ell)$ using the following statistics. Exclude the first vector z_1 from the sequence and obtain the function that minimizes the empirical risk for the remaining $\ell - 1$ elements of the sequence.

Let this function be $Q(z, \alpha_{\ell-1}|z_1)$. In this notation we indicate that the vector z_1 was excluded from the sequence (10.38). We use this excluded vector for computing the value

$$Q(z_1, \alpha_{\ell-1}|z_1).$$

Next we exclude the second vector z_2 from the sequence (the first vector is retained) and compute the value

$$Q(z_2, \alpha_{\ell-1}|z_2).$$

In this manner, we compute the values for all vectors and calculate the number of errors in the leave-one-out procedure:

$$\mathcal{L}(z_1, \dots, z_\ell) = \sum_{i=1}^{\ell} Q(z_i, \alpha_{\ell-1}|z_i).$$

We use $\mathcal{L}(z_1, \dots, z_\ell)$ as an estimate for the expectation of the function $Q(z, \alpha_\ell)$ that minimizes the empirical risk (10.37):

$$R(\alpha_\ell) = \int Q(z, \alpha_\ell) dP(z).$$

The estimator $\mathcal{L}(z_1, \dots, z_\ell)$ of the expectation is called the leave-one-out estimator.

The following theorem is valid.

Theorem 10.8 (Luntz and Brailovsky). The leave-one-out estimator is almost unbiased; that is,

$$E \frac{\mathcal{L}(z_1, \dots, z_{\ell+1})}{\ell + 1} = ER(\alpha_\ell).$$

Proof. The proof consists of the following chain of transformations:

$$\begin{aligned} E \frac{\mathcal{L}(z_1, \dots, z_{\ell+1})}{\ell + 1} &= \int \frac{1}{\ell + 1} \sum_{i=1}^{\ell+1} Q(z_i, \alpha_\ell|z_i) dP(z_1) \cdots dP(z_{\ell+1}) \\ &= \int \frac{1}{\ell + 1} \sum_{i=1}^{\ell+1} \left(\int Q(z_i, \alpha_\ell|z_i) dP(z_i) \right) \\ &\quad dP(z_1) \cdots dP(z_{i-1}) dP(z_{i+1}) \cdots dP(z_\ell) \\ &= E \frac{1}{\ell + 1} \sum_{i=1}^{\ell+1} R(\alpha_\ell|z_i) = ER(\alpha_\ell). \end{aligned}$$

The theorem is proved.

10.4.4 Proof of Theorem 10.5 and Theorem 9.2

Proof of Theorem 70.5. To prove this theorem we show that the number of errors by the leave-one-out method does not exceed the number of essential support vectors.

Indeed if the vector x_i is not an essential support vector, then there exists an expansion of the vector ψ_0 that defines the optimal hyperplane that does not contain the vector x_i .

Since the optimal hyperplane is unique, removing this vector from the training set does not change it. Therefore in the leave-one-out method it will be recognized correctly.

Thus the leave-one-out method recognizes correctly all the vectors that do not belong to the set of essential support vectors. Therefore the number $\mathcal{L}(z_1, \dots, z_{\ell+1})$ of errors in the leave-one-out method does not exceed $\mathcal{K}_{\ell+1}$ the number of essential support vectors; that is,

$$\mathcal{L}(z_1, \dots, z_{\ell+1}) \leq \mathcal{K}_{\ell+1} \quad (10.39)$$

To prove the theorem we use the result of Theorem 10.8, where we take into account the inequality (10.39):

$$ER(\alpha_\ell) = \frac{E\mathcal{L}(z_1, \dots, z_{\ell+1})}{\ell + 1} \leq \frac{E\mathcal{K}_{\ell+1}}{\ell + 1}$$

The theorem has been proved.

Proof of Theorem 9.2. To prove Theorem 9.2, note that since the number of errors in the leave-one-out procedure does not exceed the number of corrections M that makes a perceptron, the bound in Theorem 9.2 follows from the bound for M (given by Novikoff theorem) and Theorem 10.8.

10.4.5 Proof of Theorem 10.6

To prove this theorem we use another way to bound the number of errors of the leave-one-out estimator for the optimal hyperplanes.

Suppose we are given the training set

$$(x_1, y_1), \dots, (x_{\ell+1}, y_{\ell+1})$$

and the maximum of $W(a)$ in the area $a \geq 0$ is achieved at the vector $a^* = (\alpha_1^*, \dots, \alpha_\ell^*)$. Let the vector

$$\psi_0 = \sum_{i=1}^a \alpha_i^* x_i y_i$$

define the optimal hyperplane passing through the origin, where we enumerate the support vectors with $i = 1, \dots, a$.

Let us denote by a^* the vector providing the maximum for the functional $W(a)$ under constraints

$$\begin{aligned} \alpha_p &= 0, \\ \alpha_i &\geq 0 \quad (i \neq p). \end{aligned} \quad (10.40)$$

Let the vector

$$\psi_p = \sum_{i=1}^a \alpha_i^p x_i y_i$$

define the coefficients of the corresponding separating hyperplane passing through the origin.

Now denote by W_0^p the value of functional $W(\alpha)$ for

$$\begin{aligned} \alpha_i &= \alpha_i^0 & (i \neq p), \\ \alpha_p &= 0. \end{aligned} \quad (10.41)$$

Consider the vector α^p that maximizes the function $W(\alpha)$ under the constraints (10.39). The following obvious inequality is valid:

$$W(\alpha^p) \geq W_0^p$$

On the other hand the following inequality is true:

$$W(\alpha^p) \leq W(\alpha^0).$$

Therefore the inequality

$$W(\alpha^0) - W(\alpha^p) \leq W(\alpha^0) - W_0^p \quad (10.42)$$

is valid.

Now let us rewrite the right-hand side of the inequality (10.42) in the explicit form

$$\begin{aligned} W(\alpha^0) - W_0^p &= \sum_{i=1}^a \alpha_i^0 - \frac{1}{2} (\psi_0 * \psi_0) \\ &\quad - \left(\sum_{i=1}^a \alpha_i^0 - \alpha_p^0 - \frac{1}{2} ((\psi_0 - \alpha_p^0 y_p x_p) * (\psi_0 - \alpha_p^0 y_p x_p)) \right) \\ &\quad - \alpha_p^0 - \alpha_p^0 y_p (x_p * \psi_0) + \frac{1}{2} (\alpha_p^0)^2 |x_p|^2. \end{aligned}$$

Taking into account that x_p is a support vector, we have

$$W(\alpha^0) - W_0^p = \frac{1}{2} (\alpha_p^0)^2 |x_p|^2. \quad (10.43)$$

Suppose the optimal hyperplane passing through the origin recognizes the vector x_p incorrectly. This means that the inequality

$$y_p (x_p * \psi_p) \leq 0 \quad (10.44)$$

is valid. Note that as was shown in the proof of Theorem 10.5, this is possible only if the vector x_p is an essential support vector. Now let us investigate the left-hand side of the inequality (10.42). Let us fix all parameters except one; we fix α_i , $i \neq p$, and let us make one step in maximization of the function $W(\alpha)$ by changing only one parameter $a_i > 0$. We obtain

$$W(\alpha) = W(\alpha^p) + \alpha_p(1 - y_p(x_p * \psi_p)) - \frac{1}{2}\alpha_p^2|x_p|^2.$$

From this equality we obtain the best value of α_p :

$$\alpha_p = \frac{1 - y_p(x_p * \psi_p)}{|x_p|^2}$$

Increment of the function $W(\alpha)$ at this step equals

$$\Delta W_p = \frac{1}{2} \frac{(1 - y_p(x_p * \psi_p))^2}{|x_p|^2}$$

Since ΔW_p does not exceed the increment of the function $W(\alpha)$ for complete maximization, we obtain

$$W(\alpha^0) - W(\alpha^p) \geq \Delta W_p = \frac{1}{2} \frac{(1 - y_p(x_p * \psi_p))^2}{|x_p|^2}. \quad (10.45)$$

Combining (10.45), (10.43), and (10.27), we obtain

$$\frac{1}{2}(\alpha_p^0)^2|x_p|^2 \geq \frac{1}{2} \frac{(1 - y_p(x_p * \psi_p))^2}{|x_p|^2}$$

From this inequality, taking into account (10.44), we obtain

$$\alpha_p^0 \geq \frac{(1 - y_p(x_p * \psi_p))}{|x_p|^2} \geq \frac{1}{|x_p|^2}.$$

Taking into account that $|x_p| \leq \mathcal{D}_{\ell+1}$, we obtain

$$\alpha_p^0 \geq \frac{1}{\mathcal{D}_{\ell+1}^2}. \quad (10.46)$$

Thus if the optimal hyperplane makes an error classifying vector x_p in the leave-one-out procedure, then the inequality (10.45) holds. Therefore

$$\sum_{i=1}^a \alpha_i^0 \geq \frac{\mathcal{L}_{\ell+1}}{\mathcal{D}_{\ell+1}^2}, \quad (10.47)$$

where $\mathcal{L}_{\ell+1} = \mathcal{L}((x_1, y_1), \dots, (x_{\ell+1}, y_{\ell+1}))$ is the number of errors in the leave-one-out procedure on the sample $(x_1, y_1), \dots, (x_{\ell+1}, y_{\ell+1})$.

Now let us recall the properties of the optimal hyperplane (see Section 10.1)

$$(\psi_0 * \psi_0) = \sum_{i=1}^a \alpha_i^0 \quad (10.48)$$

and

$$(\psi_0 * \psi_0) = \frac{1}{\rho_{\ell+1}^2}.$$

Combining (10.46) and (10.47) with the last equation we conclude that the inequality

$$\mathcal{L}_{\ell+1} \leq \frac{\mathcal{D}_{\ell+1}^2}{\rho_{\ell+1}^2} \quad (10.49)$$

is true with probability 1.

To prove the theorem, it remains to utilize the results of Theorem 10.8:

$$ER(\alpha_\ell) = E \frac{\mathcal{L}_{\ell+1}}{\ell+1} \leq \frac{E \frac{\mathcal{D}_{\ell+1}^2}{\rho_{\ell+1}^2}}{\ell+1}.$$

The theorem has been proved.

10.4.6 Proof of Theorem 10.7

To prove the theorem it is sufficient to note that since inequalities (10.39) and (10.49) are valid, the inequality

$$\mathcal{L}_{\ell+1} \leq \min \left(\mathcal{K}_{\ell+1}, \frac{\mathcal{D}_{\ell+1}^2}{\rho_{\ell+1}^2} \right)$$

holds true with probability 1. Taking the expectation of both sides of the inequality, we prove the theorem.

10.5 THE IDEA OF THE SUPPORT VECTOR MACHINE

Now let us define the support vector machine. The support vector (SV) machine implements the following idea: It maps the input vectors \mathbf{x} into the high-dimensional *feature space* Z through some nonlinear mapping, chosen a priori. In this space, an optimal separating hyperplane is constructed (Fig. 10.2).

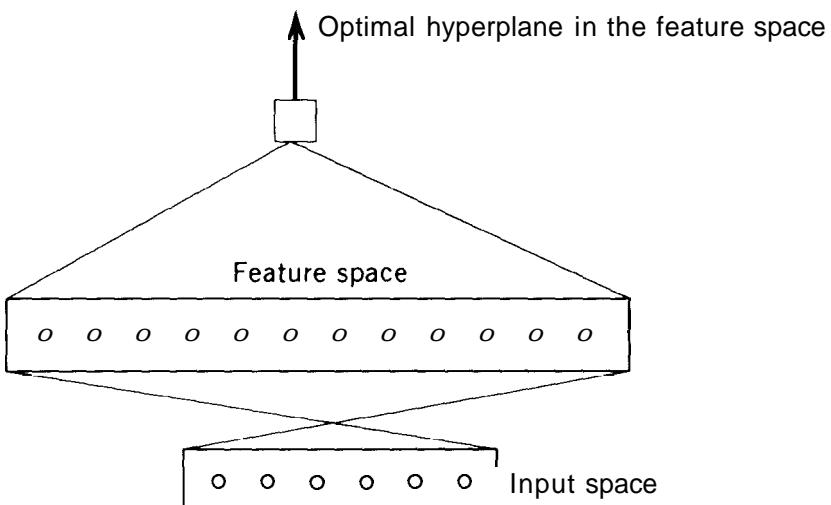


FIGURE 10.2. The SV machine maps the input space into a high-dimensional feature space and then constructs an optimal hyperplane in the feature space.

Example. To construct a decision surface corresponding to a polynomial of degree two, one can create a feature space Z that has $N = \frac{n(n+3)}{2}$ coordinates of the form

$$\begin{aligned} z^1 &= x^1, \dots, z^n = x^n && (\text{n coordinates}), \\ z^{n+1} &= (x^1)^2, \dots, z^{2n} = (x^n)^2 && (\text{n coordinates}), \\ z^{2n+1} &= x^1 x^2, \dots, z^N = x^n x^{n-1} && \left(\frac{n(n-1)}{2} \text{ coordinates} \right), \end{aligned}$$

where $x = (x^1, \dots, x^n)$. The separating hyperplane constructed in this space is a second-degree polynomial in the input space.

Two problems arise in the above approach: a conceptual and a technical one.

1. How to find a separating hyperplane that generalizes *well* (the conceptual problem). The dimensionality of the feature space is huge, and a hyperplane that separates the training data does not necessarily generalize well.
2. How to treat such high-dimensional spaces computationally (the technical problem). To construct a polynomial of degree 4 or 5 in a 200-dimensional space it is necessary to construct hyperplanes in a billion-dimensional feature space. How can this "curse of dimensionality" be overcome?

10.5.1 Generalization in High-Dimensional Space

The conceptual part of this problem can be solved by constructing the optimal hyperplane.

According to the theorems described in Section 10.3, if one can construct separating hyperplanes with a small expectation of $(D_{\ell+1}/\rho_{\ell+1})^2$ or with a small expectation of the number of support vectors, the generalization ability of the constructed hyperplanes is high, even if the feature space has a high dimensionality.

10.5.2 Mercer Theorem

However, even if the optimal hyperplane generalizes well and can theoretically be found, the technical problem of how to treat the high-dimensional feature space remains.

Note, however, that for constructing the optimal separating hyperplane in the feature space Z , one does not need to consider the feature space in *explicit form*. One only has to calculate the inner products between support vectors and the vectors of the feature space (see, for example, (10.28) and (10.30)).

Consider a general property of the inner product in a Hilbert space. Suppose one maps the vector $x \in R^n$ into a Hilbert space with coordinates

$$z_1(x), \dots, z_n(x), \dots$$

According to the Hilbert–Schmidt theory the inner product in a Hilbert space has an equivalent representation:

$$(z_1 * z_2) = \sum_{r=1}^{\infty} a_r z_r(x_1) z_r(x_2) \iff K(x_1, x_2), \quad a_r \geq 0, \quad (10.50)$$

where $K(x_1, x_2)$ is a symmetric function satisfying the following conditions.

Theorem (Mercer). To guarantee that a continuous symmetric function $K(u, v)$ in $L_2(C)$ has an expansion[†]

$$K(u, v) = \sum_{k=1}^{\infty} a_k z_k(u) z_k(v) \quad (10.51)$$

with positive coefficients $a_k > 0$ (i.e., $K(u, v)$ describes an inner product in some feature space), it is necessary and sufficient that the condition

$$\int_C \int_C K(u, v) g(u) g(v) du dv \geq 0$$

be valid for all $g \in L_2(C)$ (C being a compact subset of R^n).

[†]This means that the right-hand side of Eq. (10.50) converges to function $K(u, v)$ absolutely and uniformly.

The remarkable property of the structure of the inner product in Hilbert space that leads to construction of the SV machine is that for any kernel function $K(u, v)$ satisfying Mercer's condition there exists a feature space $(z_1(u), \dots, z_k(u), \dots)$ where this function generates the inner product (10.51).

10.5.3 Constructing SV Machines

Generating the inner product in a (high-dimensional) feature space allows the construction of decision functions that are nonlinear in the input space

$$f(x, a) = \text{sign} \left(\sum_{\text{support vectors}} y_i \alpha_i^0 K(x, x_i) + b \right) \quad (10.52)$$

and are equivalent to linear decision functions in the feature space $z_1(x), \dots, z_k(x)$,

$$f(x, a) = \text{sign} \left(\sum_{\text{support vectors}} y_i \alpha_i^0 \sum_{r=1}^{\infty} z_r(x_i) z_r(x) + b \right)$$

$(K(x, x_i))$ is the kernel that generates the inner product for this feature space). Therefore to construct function (10.52) one can use methods developed in Sections 10.2 and 10.3 for constructing linear separating hyperplanes where instead of the inner products defined as (x, x_i) , one uses the inner product defined by kernel $K(x, x_i)$.

1. To find the coefficients α_i in the separable case

$$y_i f(x_i, \alpha) = 1$$

it is sufficient to find the maximum of the functional

$$W(a) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j}^{\ell} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \quad (10.53)$$

subject to the constraints

$$\begin{aligned} \sum_{i=1}^{\ell} \alpha_i y_i &= 0, \\ \alpha_i &\geq 0, \quad i = 1, 2, \dots, \ell. \end{aligned} \quad (10.54)$$

2. To find the optimal soft margin solution for the nonseparable case, it is sufficient to maximize (10.53) under constraints

$$\begin{aligned} \sum_{i=1}^{\ell} \alpha_i y_i &= 0, \\ 0 \leq \alpha_i &\leq C. \end{aligned}$$

3. Finally to find the optimal solution for a given margin $p = 1/A$

$$f(x, \alpha) = \text{sign} \left(\frac{A}{\sqrt{\sum_{i,j=1}^{\ell} \alpha_i^0 \alpha_j^0 y_i y_j K(x_i, x_j)}} \sum_{i=1}^{\ell} \alpha_i^0 y_i K(x_i, x) + b \right)$$

one has to maximize the functional

$$W(\alpha) = \sum_{i=1}^{\ell} \alpha_i - A \sqrt{\sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j K(x_i, x_j)}$$

subject to constraints

$$\sum_{i=1}^{\ell} \alpha_i y_i = 0,$$

$$0 \leq \alpha_i \leq 1.$$

The learning machines that construct decision functions of these type are called *support vector (SV) machines*.[†] The scheme of SV machines is shown in Fig. 10.3.

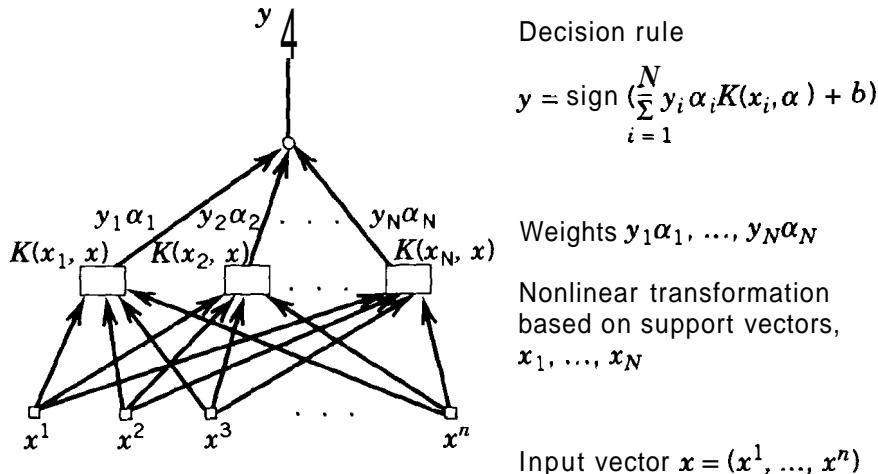


FIGURE 10.3. The support vector machine has two layers. During the learning process the first layer selects the basis $K(x, x_i)$, $i = 1, \dots, N$ (as well as the number N), from the given set of bases defined by the kernel; the second layer constructs a linear function in this space. This is completely equivalent to constructing the optimal hyperplane in the corresponding feature space.

[†] With this name we stress the idea of expanding the solution on support vectors. In SV machines the complexity of the construction depends on the number of support vectors rather than on the dimensionality of the feature space.

10.6 ONE MORE APPROACH TO THE SUPPORT VECTOR METHOD

10.6.1 Minimizing the Number of Support Vectors

The following two results of analysis of the optimal hyperplane inspire one more approach to the support vector method that is based on a linear optimization technique rather than the quadratic optimization described above[†]:

1. The optimal hyperplane has an expansion on the support vectors.
2. If a method of constructing the hyperplane has a unique solution, then the generalization ability of the constructed hyperplane depends on the number of support vectors (Theorem 10.5).

Consider the following optimization problem. Given the training data

$$(y_1, x_1), \dots, (y_\ell, x_\ell),$$

one has to find the parameters α_i , $i = 1, \dots, \ell$, and b of the hyperplane

$$\sum_{i=1}^{\ell} y_i \alpha_i (x_i * x) + b = 0, \quad \alpha_i \geq 0$$

that separates the data—that is, satisfies the inequalities

$$y_j \left(\sum_{i=1}^{\ell} y_i \alpha_i (x_i * x) + b \right) \geq 1$$

and has the smallest number of nonzero coefficients α_i .

Let us call the vector x_i that corresponds to the nonzero coefficient a , a support vector.

Let us rewrite our optimization problem in the following form: Minimize the functional

$$R = \sum_{i=1}^{\ell} \alpha_i^\sigma, \quad \alpha_i \geq 0$$

in the space of nonnegative α subject to constraints

$$y_i \left(\sum_{j=1}^{\ell} y_j \alpha_j (x_i * x_j) + b \right) \geq 1, \quad (10.55)$$

where $\sigma > 0$ is a sufficiently small value.

[†]Note that for this approach only, Theorem 10.5 is valid, while for the optimal hyperplane a more strong bound given in Theorem 10.7 holds true.

We will choose, however, $a = 1$ (the smallest value for which the solution of the optimization problem is simple). Therefore we would like to minimize the functional

$$R = \sum_{i=1}^{\ell} \alpha_i, \quad \alpha_i \geq 0 \quad (10.56)$$

subject to constraints (10.55).

10.6.2 Generalization for the Nonseparable Case

To construct the separating hyperplane for the nonseparable case using the linear optimization procedure we utilize the same idea of generalization that we used in Section 10.4. We minimize functional

$$L = \sum_{i=1}^{\ell} \alpha_i + C \sum_{i=1}^{\ell} \xi_i, \quad \alpha_i \geq 0, \quad \xi_i \geq 0 \quad (10.57)$$

over the nonnegative variables α_i , ξ_i and parameter b subject to the constraints

$$y_i \left(\sum_{j=1}^{\ell} \alpha_j (x_i * x_j) + b \right) \geq 1 - \xi_i, \quad i = 1, \dots, \ell.$$

10.6.3 Linear Optimization Method for SV Machines

To construct the support vector approximation from the set of decision rules

$$F(x; a, b) = \text{sign} \left[\sum_{j=1}^{\ell} y_j \alpha_j K(x, x_j) + b \right],$$

one can solve the linear optimization problem defined by the objective function

$$L = \sum_{i=1}^{\ell} \alpha_i + C \sum_{i=1}^{\ell} \xi_i$$

subject to the constraints

$$\alpha_i \geq 0, \quad \xi_i \geq 0, \quad i = 1, \dots, \ell,$$

$$y_i \left[\sum_{j=1}^{\ell} y_j \alpha_j K(x_i, x_j) + b \right] \geq 1 - \xi_i.$$

In this case the kernel $K(x_i, x_j)$ does not need to satisfy Mercer's conditions.

However, this construction of the SV machine does not possess all the nice statistical properties of machines constructed on the basis of the idea of the optimal hyperplane.

Therefore in the following we will consider only the SV machines constructed on the basis of the optimal hyperplane technique.

10.7 SELECTION OF SV MACHINE USING BOUNDS

The bounds on generalization ability obtained in Sections 10.2 and 10.3 allow us to choose a specific model of SV machine from a given set of models.

Note that the support vector machines implement the SRM principle. Indeed, let

$$z(x) = (z_1(x), \dots, z_N(x), \dots)$$

be the feature space and let $w = (w_1, \dots, w_N, \dots)$ be a vector of weights determining a hyperplane in this space. Consider a structure on the set of hyperplanes with elements S_k containing the functions satisfying the conditions

$$[D^2|w|^2] \leq k,$$

where D is the radius of the smallest sphere that contains the vectors $\Psi(x)$, and $|w|$ is the norm of the weights (we use canonical hyperplanes in the feature space with respect to the vectors $z_i = z(x_i)$, where x_i are the elements of the training data). According to Theorem 10.3 (now applied in the feature space), k gives an estimate of the VC dimension of the set of functions S_k defined on the training data.

The SV machine separates the training data

$$y_i [(z(x_i) * w) + b] \geq 1, \quad y_i = \{+1, -1\}, \quad i = 1, 2, \dots, \ell$$

without errors and has the minimal norm $|w|$. In other words, the SV machine separates the training data using functions from element S_k with the smallest estimate of the VC dimension.

Therefore one can use the following criteria for choosing the best model:

$$R(D_\ell, w_\ell) = D_\ell^2 |w_\ell|^2, \quad (10.58)$$

where both the value of $|w_\ell|$ and D_ℓ can be calculated from the training vectors.

Recall that in feature space the equality

$$\frac{1}{\rho_\ell^2} = |w_\ell|^2 = \sum_{i,\alpha}^{\ell} \alpha_i^0 \alpha_j^0 y_i y_j (z(x_i) * z(x_j)) = \sum_{i,\alpha}^{\ell} \alpha_i^0 \alpha_j^0 y_i y_j K(x_i, x_j) \quad (10.59)$$

holds true.

To define the radius of the smallest sphere D_ℓ that includes training vectors, one has to solve the following simple optimization problem: Minimize functional $(D_\ell * D_\ell)$ subject to constraints

$$\|z(x_i) - a\|^2 \leq D_\ell^2, \quad i = 1, \dots, \ell,$$

where \mathbf{x}_i , $i = 1, \dots, n$, are support vectors, and a is the center of the sphere.

Using technique of Lagrange multipliers, one can find that

$$D_\ell^2 = \sum_{i,j=1}^{\ell} \beta_i \beta_j (z(x_i) * z(x_j)),$$

where coefficients β_i , $i = 1, \dots, \ell$, are a solution to the following quadratic optimization problem: Maximize the functional

$$W^*(\beta) = \sum_{i=1}^{\ell} \beta_i (z(x_i) * z(x_i)) - \sum_{i,j=1}^{\ell} \beta_i \beta_j (z(x_i) * z(x_j)) \quad (10.60)$$

subject to constraints

$$\sum_{i=1}^{\ell} \beta_i = 1, \quad \beta_i \geq 0. \quad (10.61)$$

Using kernel representation, we can rewrite the radius of the smallest sphere in the form

$$D_\ell^2 = \sum_{i,j=1}^{\ell} \beta_i \beta_j K(x_i, x_j) \quad (10.62)$$

and functional (10.59) in the form

$$W^*(\beta) = \sum_{i=1}^{\ell} \beta_i K(x_i, x_i) - \sum_{i,j=1}^{\ell} \beta_i \beta_j K(x_i, x_j). \quad (10.63)$$

Therefore, to control the generalization ability of the machine (to minimize the expectation of test error), one has to construct the separating hyperplane

that minimizes the functional

$$R(D_\ell, w_\ell) = D_\ell^2 |w_\ell|^2 = \frac{D_\ell^2}{\rho_\ell^2} \quad (10.64)$$

where $|w_\ell|^2$ and D_ℓ^2 are defined by (10.59) and (10.62).

Choosing among different models (different kernels) the model that minimizes the estimate (10.58), one selects the best **SV** machine.

Thus the model selection is based on the following idea: Among different **SV** machines that separate the training data, the one with the smallest VC dimension is the best. In this section in order to estimate the VC dimension we use the upper bound (10.58). In Chapter 13, which is devoted to experiments with **SV** machines in order to obtain a more accurate evaluation, we introduce a method of direct measuring of the VC dimension from experiments with the **SV** machine. As we will see, both methods of evaluating the VC dimension (based on bound (10.58) or based on the experiments with the **SV** machine) show that the machine with the smallest estimate of VC dimension is not necessarily the one that has the smallest dimensionality of feature space.

10.8 EXAMPLES OF SV MACHINES FOR PATTERN RECOGNITION

This section uses different kernels for generating the inner products $K(x, x_i)$ in order to construct learning machines with different types of nonlinear decision surfaces in the input space. We consider three types of learning machines:

1. Polynomial **SV** machines,
2. Radial basis function **SV** machines,
3. Two-layer neural network **SV** machines.

For simplicity we discuss here the case with the separable data.

10.8.1 Polynomial Support Vector Machines

To construct polynomial of degree d decision rules, one can use the following generating kernel:

$$K(x, x_i) = [(x^* x_i) + 1]^d. \quad (10.65)$$

This symmetric function, rewritten in coordinates space, describes the inner product in the feature space that contains all the products x_{i_1}, \dots, x_{i_d} up to the degree d . Using this generating kernel, one constructs a decision

function of the form

$$f(x, a) = \text{sign} \left(\sum_{\text{support vectors}} y_i \alpha_i [(x_i * x) + 1]^d + b \right),$$

which is a factorization of d-dimensional polynomials in the n-dimensional input space.

In spite of the high dimension of the feature space (polynomials of degree d in n-dimensional input space have $O(n^d)$ free parameters), the estimate of the VC dimension of the subset of polynomials that solve real-world problems on the given training data can be low (see Theorem 10.3). If the expectation of this estimate is low, then the expectation of the probability of error is small (Theorem 10.6).

Note that both the value D_ℓ and the norm of weights $|w_\ell|$ in the feature space depend on the degree of the polynomial. This gives an opportunity to choose the best degree of polynomial for the given data by minimizing the functional (10.58).

10.8.2 Radial Basis Function SV Machines

Classical radial basis function (RBF) machines use the following set of decision rules:

$$f(x) = \text{sign} \left(\sum_{i=1}^N a_i K_\gamma(|x - x_i|) - b \right), \quad (10.66)$$

where $K_\gamma(|x - x_i|)$ depends on the distance $|x - x_i|$ between two vectors. For the theory of RBF machines see Powell (1992).

The function $K_\gamma(|z|)$ is a positive definite monotonic function for any fixed γ ; it tends to zero as $|z|$ goes to infinity. The most popular function of this type is

$$K_\gamma(|x - x_i|) = \exp\{-\gamma|x - x_i|^2\}. \quad (10.67)$$

To construct the decision rule from the set (10.66), one has to estimate

- (1) The number N of the centers x_i .
- (2) The vectors x_i , describing the centers.
- (3) The values of the parameters a_i .
- (4) The value of the parameter γ .

In the classical RBF method the first three steps (determining the parameters γ and N and vectors (centers) x_i , $i = 1, \dots, N$) are based on heuristics and only the fourth step (after finding these parameters) is determined by minimizing the empirical risk functional.

It is known (see Section 8.4) that the radial basis function of type (10.66) satisfies the condition of Mercers theorem.

Therefore one can choose the function $K_\gamma(|x - x_i|)$ as a kernel generating the inner product in some feature space. Using this kernel, one can construct a **SV** radial basis function machine.

In contrast to classical RBF methods, in the **SV** technique all four types of parameters are chosen automatically:

- (1) The number N of support vectors.
- (2) The support vectors x_i , ($i = 1, \dots, N$).
- (3) The coefficients of expansion $a_i = \alpha_i y_i$.
- (4) The width parameter γ of the kernel function chosen to minimize functional (10.58).

10.8.3 Two-Layer Neural SV Machines

One can define two-layer neural networks by choosing kernels:

$$K(x, x_i) = S[(x^* x_i)],$$

where $S(u)$ is a sigmoid function. In contrast to kernels for polynomial machines or for radial basis function machines that always satisfy Mercer's condition, the sigmoid kernel

$$S[(x^* x_i)] = \frac{1}{1 + \exp\{v(x^* x_i) - c\}},$$

satisfies the Mercer condition only for some values of parameters c and v . For example if $|x| = 1, |x_i| = 1$ the parameters c and v of the sigmoid function has to satisfy the inequality[†] $c \geq v$. For these values of parameters one can construct **SV** machines implementing the rules

$$f(x, a) = \text{sign} \left\{ \sum_{i=1}^N \alpha_i S(v(x^* x_i) - c) + b \right\}.$$

Using the technique described above, the following parameters are found

[†] In this case one can introduce two $(n+1)$ -dimensional vectors: vector x^* which first n coordinates coincide with vector x and the last coordinate is equal to zero and vector x_i^* which first n coordinate coincide with vector x_i and the last coordinate is equal to $a = \sqrt{2(c - v)}$. If $c \geq v$ then a is real and one can rewrite sigmoid function in n -dimensional input space as a radial basis function in $n+1$ dimensional space $S[(x^* x_i)] = (1 + \exp\{-0.5v\|x^* - x_i^*\|^2\})^{-1}$.

automatically:

- (1) The architecture of the two layer machine, determining the number N of hidden units (the number of support vectors),
- (2) the vectors of the weights $w_i = x_i$ in the neurons of the first (hidden) layer (the support vectors), and
- (3) the vector of weights for the second layer (values of a).

Structure of Positive Definite Functions. The important question is: How rich is the set of functions satisfying the Mercer condition?

Below we formulate four classical theorems describing the structure of the kernel functions $K(x - x_i)$ which satisfy the Mercer conditions. This type of kernels is called a positive definite function.

For positive definite functions, the following elementary properties are valid:

1. Any positive function is bounded

$$F(x - x_i) \leq F(0).$$

2. If F_1, \dots, F_n are positive definite and $a_i \geq 0$ then

$$f(x - x_i) = \sum_{k=1}^n a_k F_k(x - x_i)$$

is positive definite.

3. If each F_n is positive definite then so is

$$F(x - x_i) = \lim_{n \rightarrow \infty} F_n(x - x_i)$$

4. The product of positive definite functions is positive definite function.

In 1932, Bochner described the general structure of positive definite functions.

Theorem (Bochner). If $K(x - x_i)$ is a continuous positive definite function, then there exists a bounded nondecreasing function $V(u)$ such that $K(x - x_i)$ is a Fourier-Stieltjes transform of $V(u)$, that is

$$K(x - x_i) = \int_{-\infty}^{\infty} e^{i(x-x_i)u} dV(u).$$

If function $K(x - x_i)$ satisfies this condition, then it is positive definite.

The proof of the sufficient conditions in this theorem is obvious. Indeed

$$\begin{aligned} & \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} K(x - x_i) g(x) g(x_i) dx dx_i \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\infty} e^{i(x-x_i)u} dV(u) \right\} g(x) g(x_i) dx dx_i \\ &= \int_{-\infty}^{\infty} \left| \int_{-\infty}^{\infty} g(x) e^{ixu} \right|^2 dV(u) \geq 0. \end{aligned}$$

A particular class of positive definite functions, namely functions of the type $F(u)$, where $u = |x_i - x_j|$ plays an important role in applications (e.g., in the RBF method). Note, that $F(u)$ is a one-dimensional function but $x \in \mathbb{R}^n$. The problem is to describe functions $F(u)$ that provide positive definite functions $F(|x - x_i|)$ independent on dimensionality n of vectors x . Schoenberg (1938) described the structure of such functions,

Definition 3. Let us call function $F(u)$ completely monotonic on $(0, \infty)$, provided that it is in $C^\infty(0, \infty)$ and satisfies the conditions

$$(-1)^k F^{(k)}(u) \geq 0, \quad u \in (0, \infty), \quad k = 0, 1, \dots$$

Theorem (Schoenberg). Function $F(|x - x_i|)$ is a positive definite if and only if $F(\sqrt{|x - x_i|})$ continuous and completely monotonic.

The theorem implies that function

$$f(x - x_i) = \exp\{-a|x - x_i|^p\}, \quad a > 0$$

is positive definite if and only if $0 \leq p \leq 2$.

Lastly, a useful criterion belongs to Polya (1949).

Theorem (Polya). Any real, even, continuous function $F(u)$ which is convex on $(0, \infty)$ (that is satisfies inequality $F(1/2(u_1 + u_2)) \leq 1/2[F(u_1) + F(u_2)]$) is a positive definite.

On the basis of these theorems, one can construct different positive definite functions of type $K(x - x_i)$. For more detail about positive definite functions and their generalizations, see Stewart (1976), Micchelli (1986), and Wahba (1990).

10.9 SUPPORT VECTOR METHOD FOR TRANSDUCTIVE INFERENCE

Chapter 8 introduced a new type of inference, the transductive inference, in order to improve performance on the given test set. For a class of linear indicator functions we obtained bounds on the test error that generally speaking are better than bounds on error rate for inductive inference.

This section shows that by using the standard SV technique, one can generalize the results obtained for indicator functions which are linear in input space to nonlinear indicator functions which are linear in a feature space.

We considered the following problem: given training data

$$(y_1, x_1), \dots, (y_\ell, x_\ell) \quad y \in \{-1, 1\} \quad (10.68)$$

and test data

$$x_1^*, \dots, x_k^* \quad (10.69)$$

find in the set of linear functions

$$y = (\psi * x) + b$$

a function that minimizes the number of errors on the test set. Chapter 8 showed that in the separable case, a good solution to this problem provides a classification of the test error.

$$y_1^*, \dots, y_k^* \quad (10.70)$$

such that the joint sequence

$$(y_1, x_1), \dots, (y_\ell, x_\ell), (y_1^*, x_1^*), \dots, (y_k^*, x_k^*) \quad (10.71)$$

is separated with the maximal margin.

Therefore we would like to find such classifications (10.70) of the test vectors (10.69) for which the **optimal hyperplane**

$$y = (\psi_0^* * x) + b_0$$

maximizes the margin when it separates the data (10.71), where ψ_0^* denoted the optimal hyperplane under condition that test data (10.69) are classified according to (10.70):

$$\psi_0^* = \psi_0(y_1^*, \dots, y_k^*).$$

Let us write this formally; our goal is to find such classifications y_1^*, \dots, y_k^* for which the inequalities

$$y_i[(x_i * \psi^*) + b] \geq 1, \quad i = 1, \dots, \ell \quad (10.72)$$

$$y_j^*[(x_j^* * \psi^*) + b] \geq 1, \quad i = 1, \dots, k \quad (10.73)$$

are valid and the functional

$$\Phi(\psi_0(y_1^*, \dots, y_k^*)) = \min_{\psi_0^*} \frac{1}{2} \|\psi^*\|^2 \quad (10.74)$$

attains its minima (over classifications y_1^*, \dots, y_k^*).

In a more general setting (for a nonseparable case) find such classifications y_1^*, \dots, y_k^* for which the inequalities

$$y_i[(x_i * \psi^*) + b] \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \dots, \ell \quad (10.75)$$

$$y_j^*[(x_j^* * \psi^*) + b] \geq 1 - \xi_j^*, \quad \xi_j^* \geq 0, \quad j = 1, \dots, k \quad (10.76)$$

are valid and the functional

$$\Phi(\psi_0(y_1^*, \dots, y_k^*)) = \min_{\psi^*, \xi, \xi^*} \left[\frac{1}{2} \|\psi_0^*\|^2 + C \sum_{i=1}^{\ell} \xi_i + C^* \sum_{j=1}^k \xi_j^* \right] \quad (10.77)$$

(where C and C^* are given non-negative values) attains its minima (over y_1^*, \dots, y_k^*).

Note that to solve this problem, find the optimal hyperplanes for all fixed y_1^*, \dots, y_k^* and choose the best one. As it was shown in Section 10.2, to find the dual representation the optimal hyperplane for the fixed y_1^*, \dots, y_k^*

$$f(x) = \text{sign} \left[\sum_{i=1}^{\ell} \alpha_i y_i (x^* \cdot x_i) + \sum_{j=1}^k \alpha_j^* y_j^* (x^* \cdot x_j^*) + b \right]$$

one has to maximize the functional

$$\begin{aligned} W_{y_1^*, \dots, y_k^*}(\alpha, \alpha^*) &= \sum_{i=1}^{\ell} \alpha_i + \sum_{j=1}^k \alpha_j^* \\ &\quad - \frac{1}{2} \left[\sum_{i,r=1}^{\ell} y_i y_r \alpha_i \alpha_r (x_i * x_r) + \sum_{j,r=1}^k \alpha_j^* y_j^* y_r \alpha_r^* (x_j^* * x_r^*) \right. \\ &\quad \left. + 2 \sum_{j=1}^k \sum_{r=1}^{\ell} y_j y_r^* \alpha_j \alpha_r^* (x_j * x_r^*) \right] \end{aligned}$$

subject to constraints

$$0 \leq \alpha_i \leq C,$$

$$0 \leq \alpha_j^* \leq C^*,$$

$$\sum_{i=1}^{\ell} y_i \alpha_i + \sum_{j=1}^k y_j^* \alpha_j^* = 0.$$

Since

$$\max_{\alpha, \alpha^*} W(\alpha, \alpha^*) = \Phi(y_1^*, \dots, y_k^*)$$

to find hyperplane for transductive inference, one has to find the minimax solution where the maximum is computed by solving quadratic optimization problems and the minimum is taken over all admissible classifications of the test set.

Repeating these arguments in feature space one can formulate a transductive solution which is nonlinear in input space but linear in some feature space. Find such classification y_1^*, \dots, y_k^* for which the functional

$$W(y_1^*, \dots, y_k^*) = \max_{\alpha, \alpha^*} \left\{ \sum_{i=1}^{\ell} \alpha_i + \sum_{j=1}^k \alpha_j^* - \frac{1}{2} \left[\sum_{i,r=1}^{\ell} y_i y_r \alpha_i \alpha_r K(x_i, x_r) \right. \right. \\ \left. \left. + \sum_{j,r=1}^k \alpha_j^* y_j^* y_r \alpha_r^* K(x_j^*, x_r) + 2 \sum_{j=1}^{\ell} \sum_{r=1}^k y_j y_r^* \alpha_j \alpha_r^* K(x_j, x_r^*) \right] \right\}$$

subject to constraints

$$0 \leq \alpha_i \leq C,$$

$$0 \leq \alpha_j^* \leq C^*,$$

$$\sum_{i=1}^{\ell} y_i \alpha_i + \sum_{j=1}^k y_j^* \alpha_j^* = 0.$$

attains its minima.

Generally speaking the exact solution of this minimax problem requires searching over all possible 2^k classifications of the test set. This can be done for small number of test instances (say 3–7). For a large number of test examples, one can use various heuristic procedures (e.g., by clustering of the test data and providing the same classification for the entire cluster).

Note that the same solution can be suggested to the problem of constructing a decision rule using both labeled (10.68) and unlabeled (10.69) data. Using parameters a, a^* and b obtained in transductive solution, construct the decision rule

$$y(x) = \text{sign} \left[\sum_{i=1}^{\ell} \alpha_i y_i K(x, x_i) + \sum_{j=1}^k \alpha_j^* y_j^* K(x, x_j^*) + b \right]$$

that includes information about both data sets.

10.10 MULTICLASS CLASSIFICATION

Until now we have considered only the two-class classification problem. However real world problems often require discriminating between $n > 2$ classes.

Using a two-class classification, one can construct the n-class classifier using the following procedure:

1. Construct n two-class classification rules where rule $f_k(x), k = 1, \dots, n$ separates training vectors of the class k from the other training vectors ($\text{sign}[f_k(x_i)] = 1$, if vector x_i belongs to the class k , $\text{sign}[f_k(x_i)] = -1$ otherwise).
2. Construct the n-class classifier by choosing the class corresponding to the maximal value of functions $f_k(x_i), k = 1, \dots, n$:

$$m = \operatorname{argmax}\{f_1(x_i), \dots, f_n(x_i)\}.$$

This procedure usually gives good results.

For the SV machines however one can solve the multiclass classification problem directly[†]. Suppose we are given the training data

$$x_1^1, \dots, x_{\ell_1}^1, \dots, x_1^n, \dots, x_{\ell_n}^n,$$

where the superscript k in x_i^k denotes that the vector belongs to class k .

Consider the set of linear functions

$$f_k(x) = (x * w^k) + b_k, \quad k = 1, \dots, n.$$

Our goal is to construct n functions (n pairs (w^k, b_k)) such that the rule

$$m = \operatorname{argmax}\{(x * w^1) + b_1, \dots, (x * w^n) + b_n\}$$

separates the training data without error. That is, the inequalities

$$(x_i^k * w^k) + b_k - (x_i^m * w^m) - b_m \geq 1$$

hold true for all $k = 1, \dots, n, m \neq k$ and $i = 1, \dots, \ell_k$.

If such a solution is possible we would like to choose the pairs (w^k, b_k) , $k = 1, \dots, n$ for which the functional

$$\sum_{k=1}^n (w^k * w^k)$$

is minimal.

[†]This generalization was considered by V. Blanz and V. Vapnik. Later, similar methods were proposed independently by M. Jaakkola and by C. Watkins and J. Weston.

If the training data cannot be separated without error, we minimize the functional

$$\sum_{k=1}^n (w^k * w^k) + C \sum_{k=1}^n \sum_{i=1}^{\ell_k} \xi_i^k$$

subject to constraints

$$(x_i^k * w^k) + b_k - (x_i^m * w^m) - b_m \geq 1 - \xi_i^k,$$

where

$$k = 1, \dots, n, m \neq k, i = 1, \dots, \ell_k.$$

To solve this optimization problem we use the same optimization technique with Lagrange multipliers. We obtain:

1. Function $f_k(x)$ has the following expansion on support vectors

$$f_k(x) = \sum_{m \neq k} \sum_{i=1}^{\ell_k} \alpha_i(k, m)(x * x_i^k) - \sum_{m \neq k} \sum_{j=1}^{\ell_m} \alpha_j(m, k)(x * x_j^m) + b_k.$$

2. Coefficients $\alpha_i(k, m)$, $k = 1, \dots, n, m \neq k, i = 1, \dots, \ell_k, j = 1, \dots, \ell_m$ of this expansion have to maximize the quadratic form

$$\begin{aligned} W(\alpha) = & \sum_{k=1}^n \sum_{m \neq k} \left[\sum_{i=1}^{\ell_k} \alpha_i(k, m) \right. \\ & - \frac{1}{2} \sum_{m^* \neq k} \left(\sum_{i,j=1}^{\ell_k} \alpha_i(k, m^*) \alpha_j(k, m) (x_i^k * x_j^k) \right. \\ & + \sum_{i=1}^{\ell_m} \sum_{j=1}^{\ell_m} \alpha_i(m, k) \alpha_j(m^*, k) (x_i^m * x_j^{m^*}) \\ & \left. \left. - 2 \sum_{i=1}^{\ell_k} \sum_{j=1}^{\ell_m} \alpha_i(k, m^*) \alpha_j(m, k) (x_i^k * x_j^m) \right) \right] \end{aligned}$$

subject to constraints

$$0 \leq \sum_{m \neq k} \alpha_i(k, m) \leq C,$$

$$\begin{aligned} \sum_{m \neq k} \sum_{i=1}^{\ell_k} \alpha_i(k, m) &= \sum_{m \neq k} \sum_{j=1}^{\ell_m} \alpha_j(m, k), \\ k &= 1, \dots, n. \end{aligned}$$

For $n = 2$, this solution coincides with the two-class classification solution.

For $n > 2$ one has to estimate simultaneously $l(n - 1)$ parameters $\alpha_i(k, m)$, $i = 1, \dots, \ell_k, m \neq k, k = 1, \dots, n$, where

$$\ell = \sum_{k=1}^{\ell} \ell_k$$

To construct the n -class classifier using two-class classification rules, one needs to estimate n times ℓ parameters.

As before, to construct the SV machine we only need to replace the inner product $(x_i^r * x_j^s)$ with kernel $K(x_i^r * x_j^s)$ in the corresponding equations.

10.11 REMARKS ON GENERALIZATION OF THE SV METHOD

The SV method describes a general concept of learning machine. It considers a kernel-type function approximation that has to satisfy two conditions:

1. The kernel that defines the SV machine has to satisfy Mercer's condition.
2. The hyperplane constructed in feature space has to be optimal; that is, it possesses the smallest norm of coefficients (the largest margin).

The question arises: How crucial are these conditions? Is it possible to remove them in order to construct the general kernel method of function estimation? That is, consider functions of the form

$$y = \sum_{i=1}^{\ell} \alpha_i K(x, x_i) + b$$

(where $K(x, x_i)$ does not necessarily satisfy Mercer's condition) that approximates data using other optimality functional.

1. To answer the question about kernel, note that the generalization properties of the SV machine described in theorems presented in Section 10.3, is defined by existence of the feature space where a small norm of coefficients of the canonical hyperplane is the guarantee for good generalization. Removing Mercer's condition one removes this guarantee.
2. However, it is not necessary to use the vector-coefficient norm of the canonical hyperplane as the functional for minimization. One can minimize any positive definite quadratic form. However, to minimize arbitrary quadratic forms one has to use general quadratic optimization tools.

The following shows that for any positive definite quadratic form there exists another feature space connected to the first one by linear transformation where one achieves an equivalent solution by minimizing the norm of the coefficient vector. Solving problems in this space, one enjoys the advantage of the support vector technique.

Indeed, consider the hyperplane

$$(x * \psi) + b = 0,$$

which satisfies the inequalities

$$y_i[(x_i * \psi) + b] \geq 1 - \xi_i, \quad i = 1, \dots, \ell \quad (10.78)$$

(separates the training data) and maximizes the quadratic form

$$W(\psi) = (\psi * A\psi) + C \sum_{i=1}^{\ell} \xi_i. \quad (10.79)$$

Since A is a positive definite symmetric matrix there exists the matrix

$$B = \sqrt{A}.$$

Therefore one can rewrite the objective function as follows:

$$W(\psi) = (B\psi * B\psi) + C \sum_{i=1}^{\ell} \xi_i \quad (10.80)$$

Let us denote $\phi = B\psi$ and $z_i = B^{-1}x_i$. Then the problem of minimizing functional (10.79) subject to constraint (10.77) is equivalent to the problem of minimizing the functional

$$W(\phi) = (\phi * \phi) + C \sum_{i=1}^{\ell} \xi_i \quad (10.81)$$

subject to constraint

$$y_i[(z_i * \phi) + b] \geq 1 - \xi_i, \quad i = 1, \dots, \ell. \quad (10.82)$$

That means that there exists some linear transformation of the vectors x into vectors z for which the problem of minimizing the functional (10.80) under constraint (10.78) is equivalent to minimizing functional (10.81) under constraint (10.82).

The solution of the optimization problem with objective function (10.81) leads to the support vector technique that has important computational advantages.

THE SUPPORT VECTOR METHOD FOR ESTIMATING REAL-VALUED FUNCTIONS

In this chapter the SV method introduced in Chapter 10 for estimating indicator functions is generalized to estimate real-valued functions. The key idea in this generalization is a new type of loss function, the so-called ε -insensitive loss function. Using this type of loss function, one can control a parameter that is equivalent to the margin parameter for separating hyperplanes.

This chapter first discusses some properties of the ε -insensitive loss function and its relation to the Huber robust loss-function, then shows that the same quadratic optimization technique that was used in Chapter 10 for constructing approximations to indicator functions provides an approximation to real-valued functions, and finally introduces some kernels that are useful for the approximation of real-valued functions.

At the end of this chapter we show how the SV technique can be used for solving linear operator equations with approximately defined right-hand sides. In particular, we use the SV technique for solving integral equations that form ill-posed problems.

11.1 ε -INSENSITIVE LOSS FUNCTIONS

In Chapter 1, Section 1.4, to describe the problem of estimation of the supervisor rule $F(y|x)$ in the class of real-valued functions $\{f(x, \alpha), \alpha \in A\}$ we considered a quadratic loss function

$$M(y, f(x, \alpha)) = (y - f(x, \alpha))^2. \quad (11.1)$$

Under conditions where y is the result of measuring a regression function with

normal additive noise ξ , the ERM principle provides (for this loss function) an efficient (best unbiased) estimator of the regression $f(x, \alpha_0)$.

It is known, however, that if additive noise is generated by other laws, better approximations to the regression (for the ERM principle) give estimators based on other loss functions (associated with these laws)

$$M(y, f(x, \alpha)) = L(|y - f(x, \alpha)|)$$

($L(\xi) = -\ln p(\xi)$ for the symmetric density function $p(\xi)$).

Huber (1964) developed a theory that allows finding the best strategy for choosing the loss function using only general information about the model of the noise. In particular, he showed that if one only knows that the density describing the noise is a symmetric smooth function, then the best minimax strategy for regression approximation (the best approximation for the worst possible model of noise $p(x)$) provides the loss function

$$M(y, f(x, \alpha)) = |y - f(x, \alpha)|. \quad (11.2)$$

Minimizing the empirical risk with respect to this loss function is called the ***least modulus*** method. It belongs to the so-called ***robust regression*** family. This, however, is an extreme case where one has minimal information about the unknown density. In Section 11.3 we will discuss the key theorem of robust theory that introduces a family of robust loss functions depending on how much information about the noise is available.

To construct an SV machine for real-valued functions we use a new type of loss functions, the so-called ε -insensitive loss functions:

$$M(y, f(x, \alpha)) = L(|y - f(x, \alpha)|_\varepsilon),$$

where we denote

$$|y - f(x, \alpha)|_\varepsilon = \begin{cases} 0, & \text{if } |y - f(x, \alpha)| \leq \varepsilon, \\ |y - f(x, \alpha)| - \varepsilon & \text{otherwise.} \end{cases} \quad (11.3)$$

These loss functions describe the ε -insensitive model: The loss is equal to 0 if the discrepancy between the predicted and the observed values is less than ε .

Below we consider three loss functions

1. Linear ε -insensitive loss function:

$$L(y - f(x, \alpha)) = |y - f(x, \alpha)|_\varepsilon \quad (11.4)$$

(it coincides with the robust loss function (11.2) if $\varepsilon = 0$).

2. Quadratic ε -insensitive loss function:

$$L(y - f(x, \alpha)) = |y - f(x, \alpha)|_\varepsilon^2 \quad (11.5)$$

(it coincides with quadratic loss function (11.1) if $\varepsilon = 0$).

3. Huber loss function:

$$L(|y - f(x, \alpha)|) = \begin{cases} c|y - f(x, \alpha)| - \frac{c^2}{2} & \text{for } |y - f(x, \alpha)| > c \\ \frac{1}{2}|y - f(x, \alpha)|^2 & \text{for } |y - f(x, \alpha)| \leq c. \end{cases} \quad (11.6)$$

that we will discuss in Section 11.4.

Using the same technique, one can consider any convex loss function $L(u)$. However, the above three are special: They lead to the same simple optimization task that we used for the pattern recognition problem.

In Section 11.3 we consider methods of estimating real-valued functions that minimize the empirical risk functional with the ε -insensitive loss functions. However, in the next section we discuss the robust estimation of functions and show that the linear ε -insensitive loss function also reflects the philosophy of robust estimation.

11.2 LOSS FUNCTIONS FOR ROBUST ESTIMATORS

Consider the following situation. Suppose our goal is to estimate the expectation m of the random variable ξ using i.i.d. data

$$\xi_1, \dots, \xi_\ell.$$

Suppose also that the corresponding unknown density $p_0(\xi - m_0)$ is a smooth function, symmetric with respect to the position m_0 , and has finite second moment.

It is known that in this situation the maximum likelihood estimator

$$m = \mathcal{M}(\xi_1, \dots, \xi_\ell | p_0),$$

which maximizes

$$L(m) = \sum_{i=1}^{\ell} \ln p_0(\xi_i - m),$$

is an effective estimator. This means that among all possible unbiased estimators this estimator achieves the smallest variance; or in other words,

'Estimator $\mathcal{M}(\xi_1, \dots, \xi_\ell)$ is called unbiased if

$$E\mathcal{M}(\xi_1, \dots, \xi_\ell) = m.$$

estimator $\mathcal{M}(\xi_1, \dots, \xi_\ell | p_0)$ minimizes the functional

$$V(\mathcal{M}) = \int (\mathcal{M}(\xi_1, \dots, \xi_\ell) - m)^2 dp_0(\xi_1 - m) \dots dp_0(\xi_\ell - m). \quad (11.7)$$

Suppose now that although the density $p_0(\xi - m)$ is unknown, it is known that it belongs to some admissible set of densities $p_0(\xi - m) \in \mathcal{P}$. How do we choose an estimator in this situation? Suppose, the unknown density $p_0(\xi - m)$. However, we construct our estimator that is optimal for density is $p_1(\xi - m) \in \mathcal{P}$; that is, we define the estimator $\mathcal{M}(\xi_1, \dots, \xi_\ell | p_1)$ that maximizes the functional

$$L_1(m) = \sum_{i=1}^{\ell} \ln p_1(\xi_i - m). \quad (11.8)$$

The quality of this estimator now depends on two densities: the actual one $p_0(\xi - m)$ and the one used for constructing estimator (11.8):

$$V(p_0, p_1) = \int (\mathcal{M}(\xi_1, \dots, \xi_\ell | p_1) - m)^2 dp_0(\xi_1 - m) \dots dp_0(\xi_\ell - m).$$

Huber proved that for a wide set of admissible densities \mathcal{P} there exists a saddle point of the functional $V(p_0, p_1)$. That is, for any admissible set of densities there exists a density $p_r(\xi - m)$ such that the inequalities

$$V(p, p_r) \leq V(p_r, p_r) \leq V(p_r, p) \quad (11.9)$$

hold true for any function $p(\xi - m) \in \mathcal{P}$.

Inequalities (11.9) assert that for any admissible set of densities there exists the minimax density, the so-called ***robust density***, that in the worst scenario guarantees the smallest loss.

Using the robust density, one constructs the so-called ***robust regression estimator***. The robust regression estimator is the one that minimizes the functional

$$R_h(w) = - \sum_{i=1}^{\ell} \ln p_r(y_i - f(x_i, \mathbf{a})).$$

Below we formulate the Huber theorem, which is a foundation of the theory of robust estimation.

Consider the class H of densities formed by mixtures

$$p(\xi) = (1 - \epsilon)g(\xi) + \epsilon h(\xi)$$

of a certain fixed density $g(\xi)$ and an arbitrary density $h(\xi)$ where both densities are symmetric with respect to the origin. The weights in the mixture are $1 - \epsilon$ and ϵ , respectively. For the class of these densities the following theorem is valid.

Theorem (Huber). Let $-\ln g(\xi)$ be a twice continuously differentiable function. Then the class H possesses the following robust density

$$p_r(\xi) = \begin{cases} (1 - \epsilon)g(\xi_0) \exp\{-c(\xi_0 - \xi)\} & \text{for } \xi < \xi_0 \\ (1 - \epsilon)g(\xi) & \text{for } \xi_0 \leq \xi < \xi_1 \\ (1 - \epsilon)g(\xi_1) \exp\{-c(\xi - \xi_1)\} & \text{for } \xi \geq \xi_1, \end{cases} \quad (11.10)$$

where ξ_0 and ξ_1 are endpoints of the interval $[\xi_0, \xi_1]$ on which the monotonic (due to convexity of $-\ln g(\xi)$) function

$$-\frac{d \ln g(\xi)}{d\xi} = -\frac{g'(\xi)}{g(\xi)}$$

is bounded in absolute value by a constant c determined by the normalization condition

$$1 = (1 - \epsilon) \left(\int_{\xi_0}^{\xi_1} g(\xi) d\xi + \frac{g(\xi_0) + g(\xi_1)}{c} \right).$$

This theorem allows us to construct various robust densities. In particular, if we choose for $g(\xi)$ the normal density

$$g(\xi) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{\xi^2}{2\sigma^2}\right\}$$

and consider the class H of densities

$$p(\xi) = \frac{1 - \epsilon}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{\xi^2}{2\sigma^2}\right\} + \epsilon h(\xi),$$

then according to the theorem the density

$$p_r(\xi) = \begin{cases} \frac{1 - \epsilon}{\sqrt{2\pi}\sigma} \exp\left\{\frac{c}{2\sigma^2} - \frac{c}{\sigma}|\xi|\right\} & \text{for } |\xi| > c\sigma, \\ \frac{1 - \epsilon}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{\xi^2}{2\sigma^2}\right\} & \text{for } |\xi| \leq c\sigma \end{cases} \quad (11.11)$$

will be robust in the class, where c is determined from the normalization condition

$$1 = \frac{1 - \epsilon}{\sqrt{2\pi}\sigma} \left(\int_{-c\sigma}^{c\sigma} \exp\left\{-\frac{\xi^2}{2}\right\} d\xi + \frac{2 \exp\left\{-\frac{c^2}{2}\right\}}{c} \right).$$

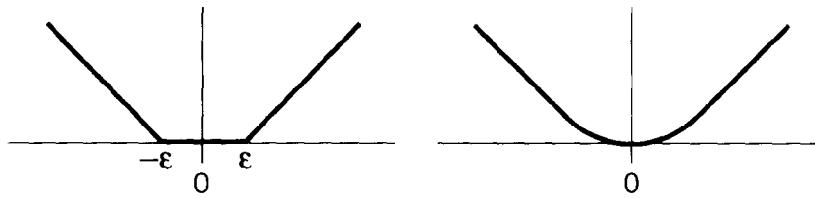


FIGURE 11.1 ϵ -insensitive linear loss function and Huber's loss function.

The loss function derived from this robust density is

$$L(\xi) = -\ln p(\xi) = \begin{cases} c|\xi| - \frac{c^2}{2} & \text{for } |\xi| > c, \\ \frac{\xi^2}{2} & \text{for } |\xi| \leq c. \end{cases} \quad (11.12)$$

It smoothly combines two functions: quadratic and linear. In one extreme case (when c tends to infinity) it defines the least-squares method, and in the other extreme case (when c tends to zero) it defines the least modulo method. In the general case, the loss functions for robust regression are combinations of two functions, one of which is $f(u) = |u|$.

Linear ϵ -insensitive loss functions, introduced in the previous section, have the same structure as robust loss functions.[†] They combine two functions; one is $f(u) = |u|$ and the other is $f(x) = 0$, which is insensitive to deviations.

It is possible to construct an SV machine for the robust loss function (11.12). However, the support vector machine defined on the basis of the linear ϵ -insensitive loss function (which has the same structure as the loss function (11.12); see Fig. 11.1) has an important advantage: In Chapter 13 we will demonstrate that by choosing the value of ϵ , one can control the number of support vectors.

11.3 MINIMIZING THE RISK WITH ϵ -INSENSITIVE LOSS FUNCTIONS

This section considers methods for constructing linear SV approximations using a given collection of data. We will obtain a solution in the form

$$f(x) = \sum_{i=1}^{\ell} \beta_i (x * x_i) + b, \quad (11.13)$$

where the coefficients β_i are nonzero only for a (small) subset of the training data (the support vectors).

[†]Formally it does not belong to the family of Huber's robust estimators since uniform distribution function does not possess a smooth derivative.

To obtain an approximation of the form (11.13) we use different loss functions that lead to different estimates of the coefficients β_i .

11.3.1 Minimizing the Risk for a Fixed Element of the Structure

Consider the structure on the set of linear functions

$$f(x, w) = \sum_{i=1}^n w_i x^i + b \quad (11.14)$$

defined in $x = (x', \dots, x^n) \in X$, where X is a bounded set in \mathbf{R}^n . Let an element S_* of the structure \mathbf{S} contain functions defined by the vector of parameters $w = (w^1, \dots, w^n)$ such that

$$(w * w) \leq A^2. \quad (11.15)$$

Suppose we are given data

$$(y_1, x_1), \dots, (x_\ell, y_\ell).$$

Our goal is to find the parameters w and b that minimize the empirical risk

$$R_{\text{emp}}(w, b) = \frac{1}{\ell} \sum_{i=1}^{\ell} |y_i - (w * x_i) - b|_{\varepsilon}^k \quad (11.16)$$

(where k is equal 1 or 2) under constraint (11.15).

This optimization problem is equivalent to the problem of finding the pair w, b that minimizes the quantity defined by slack variables $\xi_i, \xi_i^*, i = 1, \dots, \ell$

$$F(\xi, \xi^*) = \left(\sum_{i=1}^{\ell} (\xi_i^*)^k + \sum_{i=1}^{\ell} (\xi_i)^k \right) \quad (11.17)$$

under constraints

$$\begin{aligned} y_i - (w * x_i) - b &\leq \varepsilon + \xi_i^*, \quad i = 1, \dots, \ell, \\ (w * x_i) + b - y_i &\leq \varepsilon + \xi_i, \quad i = 1, \dots, \ell, \\ \xi_i^* &\geq 0, \quad i = 1, \dots, \ell, \\ \xi_i &\geq 0, \quad i = 1, \dots, \ell \end{aligned} \quad (11.18)$$

and constraint (11.15).

As before, to solve the optimization problem with constraints of inequality type one has to find the saddle point of the Lagrange functional

$$\begin{aligned}
 L(w, \xi^*, \xi; \alpha^*, \alpha, \gamma, \beta, \beta^*) &= \sum_{i=1}^{\ell} ((\xi_i^*)^k + (\xi_i)^k) - \sum_{i=1}^{\ell} \alpha_i [y_i - (w * x_i) - b + \varepsilon_i + \xi_i] \\
 &\quad - \sum_{i=1}^{\ell} \alpha_i^* [(w * x_i) + b - y_i + \varepsilon_i + \xi_i^*] - \frac{\gamma}{2} (A^2 - (w * w)) - \sum_{i=1}^{\ell} (\beta_i^* \xi_i^* + \beta_i \xi_i)
 \end{aligned} \tag{11.19}$$

(the minimum is taken with respect to elements w , b , ξ_i , and ξ_i^* and the maximum with respect to Lagrange multipliers $y \geq 0$, $\alpha_i^* \geq 0$, $a_i \geq 0$, $\beta_i^* \geq 0$, and $\beta_i \geq 0$, $i = 1, \dots, \ell$).

Minimization with respect to w , b , ξ_i^* , and ξ_i implies the following conditions:

$$w = \sum_{i=1}^{\ell} \frac{\alpha_i^* - \alpha_i}{\gamma} x_i, \tag{11.20}$$

$$\sum_{i=1}^{\ell} \alpha_i^* = \sum_{i=1}^{\ell} \alpha_i, \tag{11.21}$$

$$\begin{aligned}
 \beta_i + \alpha_i^* &\leq k(\xi_i^*)^{k-1}, & i = 1, \dots, \ell, \\
 \beta_i + \alpha_i &\leq k\xi_i^{k-1}, & i = 1, \dots, \ell.
 \end{aligned} \tag{11.22}$$

Condition (11.20) means that the desired vector w has an expansion on some elements of the training data. To find the saddle point parameters α_i^* , a_i of functional (11.19) we put (11.20) into the Lagrangian (11.19). Then taking into account (11.21) and (11.22) we determine that to find parameters α_i^* , a_i of the saddle point we have to solve the following optimization problems.

Case $k = 1$. If we consider the linear ε -insensitive loss functions, then we have to maximize the functional

$$\begin{aligned}
 W(\alpha, a^*, y) &= - \sum_{i=1}^{\ell} \varepsilon_i (\alpha_i^* + a_i) + \sum_{i=1}^{\ell} y_i (\alpha_i^* - a_i) \\
 &\quad - \frac{1}{2\gamma} \sum_{i,j=1}^{\ell} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)(x_i * x_j) - \frac{A^2 \gamma}{2}
 \end{aligned} \tag{11.23}$$

subject to constraints (11.21) and (11.22) and the constraint[†]

$$\begin{aligned} \gamma &\geq 0, \\ 0 \leq \alpha_i^*, \alpha_i &\leq 1. \end{aligned} \quad (11.24)$$

Maximizing (11.23) with respect to y , one obtains

$$\gamma = \frac{\sqrt{\sum_{i,j}^{\ell} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)(x_i * x_j)}}{A} \quad (11.25)$$

Putting this expression back in functional (11.23), we determine that to find the solution one has to maximize the functional:

$$\begin{aligned} W(\alpha, \alpha^*, \gamma) = & - \sum_{i=1}^{\ell} \varepsilon_i (\alpha_i^* + \alpha_i) + \sum_{i=1}^{\ell} y_i (\alpha_i^* - \alpha_i) \\ & - A \sqrt{\sum_{i,j=1}^{\ell} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)(x_i * x_j)} \end{aligned}$$

subject to constraints (11.21) and (11.24).

As in the pattern recognition problem, here only some of the parameters

$$\beta_i = \frac{\alpha_i^* - \alpha_i}{\gamma}, \quad i = 1, \dots, \ell$$

differ from zero. They define the support vectors of the problem. To find parameter b , it remains to minimize the empirical risk functional (11.16) with respect to b .

Case k = 2. If we consider the quadratic ε -insensitive loss function, then to find the parameters of the expansion we have to maximize the functional

$$\begin{aligned} W(\alpha, \alpha^*, \gamma) = & - \sum_{i=1}^{\ell} \varepsilon_i (\alpha_i^* + \alpha_i) + \sum_{i=1}^{\ell} y_i (\alpha_i^* - \alpha_i) \\ & - \frac{1}{2\gamma} \sum_{i,j=1}^{\ell} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)(x_i * x_j) - \frac{1}{4} \sum_{i=1}^{\ell} ((\alpha_i^*)^2 + \alpha_i^2) - \frac{A^2 \gamma}{2} \end{aligned} \quad (11.26)$$

subject to constraints (11.21), (11.24), and $y > 0$.

[†]One can solve this optimization problem using a quadratic optimization technique and line search with respect to parameter y .

Maximizing (11.26) with respect to y , we obtain that the optimal y has to satisfy the expression (11.25). Putting the expression for the optimal y back into (11.26), we obtain the functional

$$W(\alpha, \alpha^*, \gamma) = - \sum_{i=1}^{\ell} \varepsilon_i (\alpha_i^* + \alpha_i) + \sum_{i=1}^{\ell} y_i (\alpha_i^* - \alpha_i)$$

$$- A \sqrt{\sum_{i,j=1}^{\ell} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)(x_i * x_j)} - \frac{1}{4} \sum_{i=1}^{\ell} ((\alpha_i^*)^2 + \alpha_i^2),$$

which one has to maximize under constraints (11.21) and (11.24). The obtained parameters define the vector coefficients (11.20) of the desired hyperplane.

11.3.2 The Basic Solutions

One can reduce the optimization problem of finding the vector w to a quadratic optimization problem if, instead of minimizing the functional (11.17), subject to constraints (11.15) and (11.18), one minimizes the functional

$$\Phi(w, \xi^*, \xi) = \frac{1}{2}(w * w) + \frac{C}{k} \left(\sum_{i=1}^{\ell} (\xi_i^*)^k + \sum_{i=1}^{\ell} (\xi_i)^k \right)$$

(with a given value C) subject to constraints (11.18), where $k = 1$ for the linear ε -insensitive loss function and $k = 2$ for the quadratic ε -insensitive loss function.

Repeating the same arguments as in the previous section (constructing a Lagrange functional, minimizing it with respect to variables w, ξ_i, ξ_i^* , $i = 1, \dots, \ell$, and excluding these variables from the Lagrangian), one obtains that the desired vector has the following expansion:

$$w = \sum_{i=1}^{\ell} (\alpha_i^* - \alpha_i) x_i. \quad (11.27)$$

Case $k = 1$. To find coefficients α_i^* , α_i , $i = 1, \dots, \ell$, for case $k = 1$, one has to maximize the quadratic form

$$W(\alpha, \alpha^*) = - \sum_{i=1}^{\ell} \varepsilon_i (\alpha_i^* + \alpha_i) + \sum_{i=1}^{\ell} y_i (\alpha_i^* - \alpha_i)$$

$$- \frac{1}{2} \sum_{i,j=1}^{\ell} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)(x_i * x_j) \quad (11.28)$$

subject to constraints

$$\begin{aligned}\sum_{i=1}^{\ell} \alpha_i^* &= \sum_{i=1}^{\ell} \alpha_i, \\ 0 \leq \alpha_i^* &\leq C, \quad i = 1, \dots, \ell, \\ 0 \leq \alpha_i &\leq C, \quad i = 1, \dots, \ell.\end{aligned}$$

From (11.28) it is easy to see that for any $i = 1, \dots, \ell$ the equality

$$\alpha_i^* \times \alpha_i = 0$$

holds true. Therefore, for the particular case where $\varepsilon = 0$ and $y_i \in \{-1, 1\}$, the considered optimization problems coincide with those described for pattern recognition in Chapter 10, Section 10.4. We use this solution in Chapter 13 for solving real-life problems.

Case k = 2. To find the solution (coefficients of expansion α_i^*, α_i in (11.27)) for the case $k = 2$, one has to maximize the quadratic form

$$\begin{aligned}W(\alpha, \alpha^*) &= -\sum_{i=1}^{\ell} \varepsilon_i (\alpha_i + \alpha_i^*) + \sum_{i=1}^{\ell} y_i (\alpha_i^* - \alpha_i) \\ &\quad - \frac{1}{2} \left(\sum_{i,j=1}^{\ell} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)(x_i * x_j) + \frac{1}{C} \sum_{i=1}^{\ell} (\alpha_i^*)^2 + \frac{1}{C} \sum_{i=1}^{\ell} (\alpha_i)^2 \right)\end{aligned}$$

subject to constraints

$$\begin{aligned}\sum_{i=1}^{\ell} \alpha_i^* &= \sum_{i=1}^{\ell} \alpha_i, \\ \alpha_i^* \geq 0, &\quad i = 1, \dots, \ell, \\ \alpha_i \geq 0, &\quad i = 1, \dots, \ell.\end{aligned}$$

11.3.3 Solution for the Huber Loss Function

Lastly, consider the SV machine for the Huber loss function

$$F(\xi) = \begin{cases} c|\xi| - \frac{c^2}{2} & \text{for } |\xi| \leq c, \\ \frac{1}{2}\xi^2 & \text{for } |\xi| > c. \end{cases}$$

Let us minimize the functional

$$\Phi(w, \xi^*, \xi) = \frac{1}{2}(w * w) + C \left(\sum_{i=1}^{\ell} F(\xi_i^*) + \sum_{i=1}^{\ell} F(\xi_i) \right)$$

subject to constraints

$$\begin{aligned} y_i - (w * x_i) - b &\leq \xi_i^*, \quad i = 1, \dots, \ell, \\ (w * x_i) + b - y_i &\leq \xi_i, \quad i = 1, \dots, \ell, \\ \xi_i^* &\geq 0, \quad i = 1, \dots, \ell, \\ \xi_i &\geq 0, \quad i = 1, \dots, \ell. \end{aligned}$$

For this loss function, to find the desired linear function

$$(w_0 * x) + b = \sum_{i=1}^{\ell} (\alpha_i^* - \alpha_i)(x_i * x) + b,$$

one has to find the coefficients α_i^* and α_i that maximize the quadratic form

$$\begin{aligned} W(\alpha, \alpha^*) &= \sum_{i=1}^{\ell} y_i (\alpha_i^* - \alpha_i) \\ &- \frac{1}{2} \left(\sum_{i,j=1}^{\ell} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)(x_i * x_j) + \frac{c}{C} \sum_{i=1}^{\ell} (\alpha_i^*)^2 + \frac{c}{C} \sum_{i=1}^{\ell} (\alpha_i)^2 \right) \end{aligned}$$

subject to constraints

$$\begin{aligned} \sum_{i=1}^{\ell} \alpha_i^* &= \sum_{i=1}^{\ell} \alpha_i, \\ 0 \leq \alpha_i^* &\leq C, \quad i = 1, \dots, \ell. \end{aligned}$$

When $c = \varepsilon < 1$ the solution obtained for the Huber loss function is close to the solution obtained for the ε -insensitive loss function. However, the expansion of the solution for the ε -insensitive loss function uses fewer support vectors.

11.4 SV MACHINES FOR FUNCTION ESTIMATION

Now we are ready to construct the support vector machine for real-valued function estimation problems. As in the pattern recognition case, we map the input vectors x into high-dimensional feature space Z where we consider linear functions

$$f(x, \beta) = (z * w) + b = \sum_{i=1}^{\ell} \beta_i (z * z_i) + b. \quad (11.29)$$

As in the pattern recognition case we will not perform the mapping explicitly. We will perform it implicitly by using kernels for estimating the inner product in feature space. To construct the linear function (11.29) in feature space Z we use results obtained in the previous section with only one correction: In all formulas obtained in Section 11.3 we replace the inner product in input space $(x_i * x_j)$ with the inner product in feature space described by the corresponding kernel $K(x_i, x_j)$ satisfying Mercer's condition. (For kernel representation of inner products in feature space see Chapter 10, Section 10.5.)

Therefore our linear function in feature space (11.29) has the following equivalent representation in input space:

$$f(x, \beta) = \sum_{i=1}^{\ell} \beta_i K(x, x_i) + b, \quad (11.30)$$

where β_i , $i = 1, \dots, \ell$, are scalars; x_i , $i = 1, \dots, \ell$, are vectors; and $K(x, x_i)$ is a given function satisfying Mercer's conditions.

To find functions of form (11.30) that are equivalent (in feature space) to the function (11.29) we use the same optimization methods that were used in Section 11.3.

11.4.1 Minimizing the Risk for a Fixed Element of the Structure in Feature Space

As in Section 11.3.1, consider the structure on the set of linear functions defined by the norm of coefficients of linear functions in a feature space:

$$(w * w) \leq A^2. \quad (11.31)$$

Suppose that we are given the observations

$$(y_1, x_1), \dots, (y_\ell, x_\ell),$$

which in the feature space are

$$(y_1, z_1), \dots, (y_\ell, z_\ell).$$

To find an approximation of the form (11.30) that is equivalent to a linear function minimizing the empirical risk functional in feature space

$$R_{\text{emp}}(w, b) = \frac{1}{\ell} \sum_{i=1}^{\ell} |y_i - (w * z_i) - b|_{\epsilon_i}^k,$$

subject to constraint (11.31), one has to find coefficients

$$\beta_i = \frac{\alpha_i^* - \alpha_i}{\gamma}, \quad i = 1, \dots, \ell$$

of expansion

$$\beta = \sum_{i=1}^{\ell} \beta_i z_i,$$

where α_i^* , a , and y are the parameters that maximize the following functionals:

Case k = 1. For the linear ε -insensitive loss function one has to maximize the functional

$$W(\alpha, \alpha^*, \gamma) = - \sum_{i=1}^{\ell} \varepsilon_i (\alpha_i^* + \alpha_i) + \sum_{i=1}^{\ell} y_i (\alpha_i^* - \alpha_i) \\ - A \sqrt{\sum_{i,j=1}^{\ell} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) K(x_i, x_j)} \quad (11.32)$$

subject to the constraint

$$\sum_{i=1}^{\ell} \alpha_i^* = \sum_{i=1}^{\ell} \alpha_i \quad (11.33)$$

and to the constraints

$$0 \leq \alpha_i^* \leq 1, \quad 0 \leq \alpha_i \leq 1, \quad i = 1, \dots, \ell. \quad (11.34)$$

Case k = 2. For the quadratic ε -insensitive loss function, one has to maximize the functional

$$W(\alpha, \alpha^*, \gamma) = - \sum_{i=1}^{\ell} \varepsilon_i (\alpha_i^* + \alpha_i) + \sum_{i=1}^{\ell} y_i (\alpha_i^* - \alpha_i) \\ - A \sqrt{\sum_{i,j=1}^{\ell} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) K(x_i, x_j)} - \frac{1}{4} \sum_{i=1}^{\ell} [(\alpha_i^*)^2 + \alpha_i^2] \quad (11.35)$$

subject to constraints (11.33) and (11.34). (Compare to results of Section 4.1.)

11.4.2 The Basic Solutions in Feature Space

To find function (11.30) that is equivalent to one that minimizes the functional

$$\Phi(w, \xi^*, \xi) = \frac{1}{2}(w * w) + \frac{C}{2} \left(\sum_{i=1}^{\ell} (\xi_i^*)^k + \sum_{i=1}^{\ell} (\xi_i)^k \right), \quad k = 1, 2 \quad (11.36)$$

subject to constraints

$$|y_i - (w * z_i)| \leq \varepsilon_i + \xi_i, \quad i = 1, \dots, \ell, \quad (11.37)$$

one has to find

$$\beta_i = \alpha_i^* - \alpha_i, \quad i = 1, \dots, \ell,$$

where the parameters are such that:

Case k = 1. Parameters α_i^* and α_i maximize the function

$$\begin{aligned} W(\alpha, \alpha^*) = & - \sum_{i=1}^{\ell} \varepsilon_i (\alpha_i^* + \alpha_i) + \sum_{i=1}^{\ell} y_i (\alpha_i^* - \alpha_i) \\ & - \frac{1}{2} \sum_{i,j=1}^{\ell} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) K(x_i, x_j) \end{aligned}$$

subject to the constraint

$$\sum_{i=1}^{\ell} \alpha_i^* = \sum_{i=1}^{\ell} \alpha_i$$

and to the constraints

$$\begin{aligned} 0 \leq \alpha_i^* & \leq C, \quad i = 1, \dots, \ell, \\ 0 \leq \alpha_i & \leq C, \quad i = 1, \dots, \ell. \end{aligned}$$

Case k = 2. Parameters α_i^* and α_i maximize the quadratic form

$$\begin{aligned} W(\alpha, \alpha^*) = & -\varepsilon \sum_{i=1}^{\ell} (\alpha_i + \alpha_i^*) + \sum_{i=1}^{\ell} y_i (\alpha_i^* - \alpha_i) \\ & - \frac{1}{2} \left(\sum_{i,j=1}^{\ell} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) K(x_i, x_j) + \frac{1}{C} \sum_{i=1}^{\ell} (\alpha_i^*)^2 + \frac{1}{C} \sum_{i=1}^{\ell} (\alpha_i)^2 \right) \end{aligned}$$

subject to constraints

$$\begin{aligned} \sum_{i=1}^{\ell} \alpha_i^* & = \sum_{i=1}^{\ell} \alpha_i, \\ \alpha_i^* & \geq 0, \quad i = 1, \dots, \ell, \\ \alpha_i & \geq 0, \quad i = 1, \dots, \ell. \end{aligned}$$

When $\varepsilon = 0$ and

$$K(x_i, x_j) = \text{Cov}\{f(x_i), f(x_j)\}$$

is a covariance function of a stochastic process with

$$Ef(x) = 0,$$

the obtained solution coincides with the so-called *krieging* method developed in geostatistics (see Matheron, 1973).

11.4.3 Solution for Huber Loss Function in Feature Space

To minimize functional

$$\Phi(w, \xi^*, \xi) = \frac{1}{2}(w * w) + C \left(\sum_{i=1}^{\ell} F(\xi_i^*) + \sum_{i=1}^{\ell} F(\xi_i) \right)$$

subject to constraints

$$\begin{aligned} y_i - (w * z_i) - b &\leq \xi_i^*, \quad i = 1, \dots, \ell, \\ (w * z_i) + b - y_i &\leq \xi_i, \quad i = 1, \dots, \ell, \\ \xi_i^* &\geq 0, \quad i = 1, \dots, \ell, \\ \xi_i &\geq 0, \quad i = 1, \dots, \ell \end{aligned}$$

with the Huber loss function

$$F(\xi) = \begin{cases} c|\xi| - \frac{c^2}{2} & \text{for } |\xi| \leq c, \\ \frac{1}{2}\xi^2 & \text{for } |\xi| > c, \end{cases}$$

one has to find the parameters $\beta = \alpha_i^* - \alpha_i$, $i = 1, \dots, \ell$, that maximize the functional

$$\begin{aligned} W(\alpha, \alpha^*) &= \sum_{i=1}^{\ell} y_i (\alpha_i^* - \alpha_i) \\ &- \frac{1}{2} \left(\sum_{i,j=1}^{\ell} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) K(x_i, x_j) + \frac{c}{C} \sum_{i=1}^{\ell} (\alpha_i^*)^2 + \frac{c}{C} \sum_{i=1}^{\ell} (\alpha_i)^2 \right) \end{aligned}$$

subject to constraints

$$\begin{aligned} \sum_{i=1}^{\ell} \alpha_i^* &= \sum_{i=1}^{\ell} \alpha_i \\ 0 \leq \alpha, \alpha_i^* &\leq C, \quad i = 1, \dots, \ell. \end{aligned}$$

11.4.4 Linear Optimization Method

As in the pattern recognition case, one can simplify the optimization problem even more by reducing it to a linear optimization task. Suppose we are given data

$$(y_1, x_1), \dots, (y_\ell, x_\ell).$$

Let us approximate functions using functions from the set

$$y(x) = \sum_{i=1}^{\ell} \beta_i K(x_i, x) + b,$$

where β_i is some real value, x_i is a vector from a training set, and $K(x_i, x)$ is a kernel function. We call the vectors from the training set that correspond to nonzero β_i the *support vectors*. Let us rewrite β_i in the form

$$\beta_i = \alpha_i^* - a,,$$

where $\alpha_i^* > 0$, $\alpha_i > 0$.

One can use as an approximation the function that minimizes the functional

$$W(\alpha, \xi) = \sum_{i=1}^{\ell} \alpha_i + \sum_{i=1}^{\ell} \alpha_i^* + C \sum_{i=1}^{\ell} \xi_i + C \sum_{i=1}^{\ell} \xi_i^*$$

subject to constraints

$$\begin{aligned} \alpha_i &\geq 0, \quad \alpha_i^* \geq 0, \quad i = 1, \dots, \ell, \\ \xi_i &\geq 0, \quad \xi_i^* \geq 0, \\ y_i - \sum_{j=1}^{\ell} (\alpha_j^* - \alpha_j) K(x_i, x_j) - b &\leq \varepsilon - \xi_i^* \\ \sum_{j=1}^{\ell} (\alpha_j^* - \alpha_j) K(x_i, x_j) + b - y_i &\leq \varepsilon - \xi_i. \end{aligned}$$

The solution to this problem requires only linear optimization techniques.

11.4.5 Multi-Kernel Decomposition of Functions

Using the linear optimization technique, one can construct a method of multi-kernel function approximation that uses data

$$(y_1, x_1), \dots, (y_\ell, x_\ell),$$

constructs the SV approximation with a small number of support vectors.

Consider p kernel-functions

$$K_1(x, x_i), \dots, K_p(x, x_i).$$

We seek a solution that has the following form

$$f(x) = \sum_{i=1}^{\ell} \sum_{m=1}^p K_m(x, x_i)(\alpha_j^*(m) - \alpha_j(m)) + b,$$

where coefficients $\alpha_j^*(m)$, $\alpha_j(m)$ and slack variables ξ_i , ξ_i^* minimize the functional

$$R = \sum_{i=1}^{\ell} \sum_{m=1}^p (\alpha_i(m) + \alpha^*(m)) + C \sum_{i=1}^{\ell} \xi_i + C \sum_{i=1}^{\ell} \xi_i^*$$

subject to constraints

$$\begin{aligned} y_i - \sum_{j=1}^{\ell} \sum_{m=1}^p K_m(x_j, x_i)(\alpha_j^*(m) - \alpha_j(m)) - b &\leq \varepsilon_i + \xi_i, \quad i = 1, \dots, \ell \\ \sum_{j=1}^{\ell} \sum_{m=1}^p K_m(x_j, x_i)(\alpha_j^*(m) - \alpha_j(m)) + b - y_i &\leq \varepsilon_i + \xi_i^*, \quad i = 1, \dots, \ell \\ \alpha_j(m) &\geq 0, \alpha_j^* \geq 0, j = 1, \dots, \ell, \quad m = 1, \dots, p \end{aligned}$$

The idea of a multi-kernel decomposition was suggested to solve the density estimation problem (see Weston et al., 1998).

11.5 CONSTRUCTING KERNELS FOR ESTIMATION OF REAL-VALUED FUNCTIONS

To construct different types of SV machines, one has to choose different kernels $K(x, x_i)$ satisfying Mercer's condition.

In particular, one can use the same kernels that were used for approximation of indicator functions:

1. Kernels generating polynomials:

$$K(x, x_i) = [(x * x_i) + 1]^d.$$

2. Kernels generating radial basis functions:

$$K(x, x_i) = K(|x - x_i|),$$

for example,

$$K(|x - x_i|) = \exp \left\{ -\gamma |x - x_i|^2 \right\}.$$

3. Kernels generating two-layer neural networks:

$$K(x, x_i) = S(v(x * x_i) + c), \quad c \geq v, \|x\| = 1.$$

On the basis of these kernels, one can obtain the approximation

$$f(x, \alpha_0) = \sum_{i=1}^{\ell} \beta_i K(x, x_i) + b \quad (11.38)$$

using the optimization techniques described above.

In the pattern recognition problem we used function (11.38) under the discrimination sign; that is, we considered functions $\text{sign}[f(x, a)]$.

However, the problem of approximation of real-valued functions is more delicate than the approximation of indicator functions (the absence of $\text{sign}\{\cdot\}$ in front of function $f(x, a)$ significantly changes the problem of approximation).

Various real-valued function estimation problems need various sets of approximating functions. Therefore it is important to construct special kernels that reflect special properties of approximating functions.

To construct such kernels we will use two main techniques:

1. Constructing kernels for approximating one-dimensional functions and
2. Composition of multidimensional kernels using one-dimensional kernels.

11.5.1 Kernels Generating Expansion on Polynomials

To construct kernels that generate expansion of one-dimensional functions in the first N terms of orthonormal polynomials $P_i(x), i = 1, \dots, N$, one can use the following Christoffel–Darboux formula

$$\begin{aligned} K_n(x, y) &= \sum_{k=1}^n P_k(x)P_k(y) = a_n \frac{P_{n+1}(x)P_n(y) - P_n(x)P_{n+1}(y)}{x - y}, \\ K_n(x, x) &= \sum_{k=1}^n P_k^2(x) = a_n [P'_{n+1}(x)P_n(x) - P'_n(x)P_{n+1}(x)], \end{aligned} \quad (11.39)$$

where a_n is a constant that depends on the type of polynomial and the number n of elements in the orthonormal basis.

One can show that by increasing n , the kernels $K(x, y)$ approach the δ -function. Consider the kernel

$$K(x, y) = \sum_{i=1}^{\infty} r_i \psi_i(x) \psi_i(y), \quad (11.40)$$

where $r_i > 0$ converges to zero as i increases. This kernel defines regularized expansion on polynomials.

We can choose values r_i such that they improve the convergence properties of series (11.40). For example, we can choose $r_i = q^i$, $0 \leq q \leq 1$.

Example. Consider the (one-dimensional) Hermite polynomials

$$H_k(x) = \mu_k P_k(x) e^{-x^2}, \quad (11.41)$$

where

$$P_k(x) = (-1)^k e^{x^2} \left(\frac{d}{dx} \right)^k e^{-x^2}$$

and μ_k are normalization constants.

For these polynomials, one can obtain the kernels

$$\begin{aligned} K(x, y) &= \sum_{i=0}^{\infty} q^i H_i(x) H_i(y) \\ &= \frac{1}{\sqrt{\pi(1-q^2)}} \exp \left\{ \frac{2xyq}{1+q} - \frac{(x-y)^2 q^2}{1-q^2} \right\} \end{aligned} \quad (11.42)$$

(Titchmarsh, 1948; Mikhlin, 1964).

To construct our kernels we do not even need to use orthonormal bases. In the next section, we use linearly independent bases that are not orthogonal to construct kernels for spline approximations.

Such generality (any linearly independent system with any smoothing parameters) opens wide opportunities to construct one-dimensional kernels for SV machines.

11.5.2 Constructing Multidimensional Kernels

Our goal is to construct kernels for approximating multidimensional functions defined on the vector space $X \subset R^n$ where all coordinates of vector $x = (x^1, \dots, x^n)$ are defined on the same finite or infinite interval I .

Suppose now that for any coordinate x^k the complete orthonormal basis $b_{i_k}(x^k)$, $i = 1, 2, \dots$, is given. Consider the following set of basis functions:

$$b_{i_1, i_2, \dots, i_n}(x^1, \dots, x^n) = b_{i_1}(x^1) b_{i_2}(x^2) \cdot \dots \cdot b_{i_n}(x^n) \quad (11.43)$$

in the n -dimensional space. These functions are constructed from the coordinatewise basis functions by direct multiplication (tensor product) of the basis functions, where all indexes i_k take all possible integer values from 1 to ∞ .

It is known that the set of functions (11.43) is a complete orthonormal basis in $\mathbf{X} \subset \mathbb{R}^n$.

Now let us consider the more general situation where a (finite or infinite) set of coordinatewise basis functions is not necessarily orthonormal. Consider as a basis of n -dimensional space the tensor product of coordinatewise basis.

For this structure of multidimensional spaces the following theorem is true.

Theorem 11.1. Let a multidimensional set of functions be defined by the basis functions that are tensor products of coordinatewise basis functions. *Then* the kernel that defines the inner product in the n -dimensional basis is the *product* of n one-dimensional kernels.

Proof Consider two vectors $x = (x^1, \dots, x^n)$ and $y = (y^1, \dots, y^n)$ in n -dimensional space. According to the definition the kernel describing the inner product for these two vectors in the feature space is

$$\begin{aligned} K(x, y) &= \sum_{i_1, \dots, i_n} b_{i_1, \dots, i_n}(x^1, \dots, x^n) b_{i_1, \dots, i_n}(y^1, \dots, y^n) \\ &= \sum_{i_1, \dots, i_n} b_{i_1, \dots, i_{n-1}}(x^1, \dots, x^{n-1}) b_{i_1, \dots, i_{n-1}}(y^1, \dots, y^{n-1}) b_{i_n}(x^n) b_{i_n}(y^n) \\ &= \sum_{i_1, \dots, i_{n-1}} b_{i_1, \dots, i_{n-1}}(x^1, \dots, x^{n-1}) b_{i_1, \dots, i_{n-1}}(y^1, \dots, y^{n-1}) \sum_{i_n} b_{i_n}(x^n) b_{i_n}(y^n) \\ &= K_1(x^n, y^n) \sum_{i_1, \dots, i_{n-1}} b_{i_1, \dots, i_{n-1}}(x^1, \dots, x^{n-1}) b_{i_1, \dots, i_{n-1}}(y^1, \dots, y^{n-1}). \end{aligned}$$

Reiterating this convolution, we obtain

$$K(x, y) = \prod_{k=1}^n K_k(x^k, y^k). \quad (11.44)$$

The theorem has been proved.

Continuation of Example. Now let us construct a kernel for the regularized expansion on n -dimensional Hermite polynomials. In the example discussed above we constructed a kernel for one dimensional Hermite polynomials. According to Theorem 11.1 if we consider as a basis of n -dimensional space the tensor product of one dimensional basis-functions then the kernel for generating n -dimensional expansion is the product of n one-dimensional kernels

$$\begin{aligned} K(x, y) &= \prod_{i=1}^n \frac{1}{\sqrt{\pi(1-q^2)}} \exp \left\{ \frac{2x^i y^i q}{1+q} - \frac{(x^i - y^i)^2 q^2}{1-q^2} \right\} \\ &\quad - \frac{1}{(1-q^2)^{n/2}} \exp \left\{ \frac{2(x * y)q}{1+q} - \frac{|x-y|^2 q^2}{1-q^2} \right\}. \end{aligned} \quad (11.45)$$

Thus, we obtained a kernel for constructing semilocal approximations:

$$K(x, y) = C \exp\{2(x * y)\delta\} \exp\{-|x - y|^2 \sigma^2\}, \quad \delta, \sigma > 0, \quad (11.46)$$

where the multiplier with the inner product of two vectors defines "global" approximation since the Gaussian defines the vicinity of approximation (compare to the result of Chapter 6, Section 6.6 for local function approximation).

11.6 KERNELS GENERATING SPLINES

Below we introduce the kernels that can be used to construct a spline approximation of high-dimensional functions. We will construct splines with both a fixed number of knots and an infinite number of knots. In all cases the computational complexity of the solution depends on the number of support vectors that one needs to approximate the desired function with ϵ -accuracy, rather than on the dimensionality of the space or on the number of knots.

11.6.1 Spline of Order d with a Finite Number of Knots

Let us start by describing the kernel for approximation of one-dimensional functions on the interval $[0, a]$ by splines of order $d \geq 0$ with m knots:

$$(t_1, \dots, t_m), \quad t_i = \frac{ia}{m}, \quad i = 1, \dots, m.$$

According to the definition, spline approximations have the form (Fig 11.2)

$$f(x) = \sum_{r=0}^d a_r^* x^r + \sum_{i=1}^m a_i (x - t_i)_+^d. \quad (11.47)$$

Consider the following mapping of the one-dimensional variable x into an $(m+d+1)$ -dimensional vector u :

$$x \longrightarrow u = (1, x, \dots, x^d, (x - t_1)_+^d, \dots, (x - t_m)_+^d),$$

where we denote

$$(x - t_k)_+^d = \begin{cases} 0 & \text{if } x \leq t_k, \\ (x - t_k)^d & \text{if } x > t_k. \end{cases}$$

Since spline function (11.47) can be considered as the inner product of two vectors

$$f(x) = (a^* u)$$

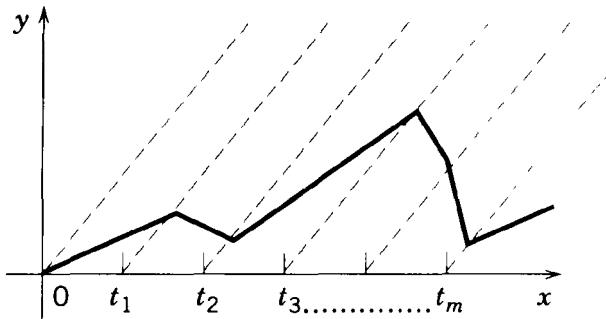


FIGURE 11.2. Using an expansion on the functions $\mathbf{1}, x, (x - t_1)_+, \dots, (x - t_m)_+$, one can construct a piecewise linear approximation of a function. Analogously, an expansion on the functions $1, x, \dots, x^d, (x - t_1)_+^d, \dots, (x - t_m)_+^d$ provides piecewise polynomial approximation.

(where $\mathbf{a} = (a_0, \dots, a_{m+d})$), one can define the kernel that generates the inner product in feature space as follows:

$$K(x, x_t) = (u * u_t) = \sum_{r=0}^d x^r x_t^r + \sum_{i=1}^m (x - t_i)_+^d (x_t - t_i)_+^d. \quad (11.48)$$

Using the generating kernel (11.48), the SV machine constructs the function

$$f(x, \beta) = \sum_{i=1}^{\ell} \beta_i K(x, x_i) + b,$$

that is, a spline of order d defined on m knots.

To construct kernels generating splines in n -dimensional spaces, note that n -dimensional splines are defined as an expansion on the basis functions that are tensor products of one dimensional basis functions. Therefore according to the Theorem 11.1, kernels generating n -dimensional splines are the product of n one-dimensional kernels:

$$K(x, x_i) = \prod_{k=1}^n K(x^k, x_i^k),$$

where we denoted $x = (x^1, \dots, x^n)$.

11.6.2 Kernels Generating Splines with an Infinite Number of Knots

In applications of SV machines the number of knots does not play an important role (the values of ε_i are more important). Therefore to simplify the

calculation, we use splines with an infinite number of knots defined on the interval $(0, a)$, $0 < a < \infty$, as the expansion

$$f(x) = \sum_{i=0}^d a_i x^i + \int_0^a a(t)(x - t)_+^d dt,$$

where a_i , $i = 0, \dots, d$, are unknown values and $a(t)$ is an unknown function that defines the expansion. One can consider this expansion as an inner product. Therefore one can construct the following kernel for generating splines of order d with an infinite number of knots:

$$\begin{aligned} K(x_j, x_i) &= \int_0^a (x_j - t)_+^d (x_i - t)_+^d dt + \sum_{r=0}^d x_j^r x_i^r \\ &= \int_0^{(x_j \wedge x_i)} (x_j - t)^d (x_i - t)^d dt + \sum_{r=0}^d x_j^r x_i^r \\ &= \int_0^{(x_j \wedge x_i)} u^d (u + |x_j - x_i|)^d du + \sum_{r=0}^d x_j^r x_i^r \\ &= \sum_{r=0}^d \frac{C_d^r}{2d - r + 1} (x_j \wedge x_i)^{2d-r+1} |x_j - x_i|^r + \sum_{r=0}^d x_j^r x_i^r, \quad (11.49) \end{aligned}$$

where we denote $\min(x, x_i) = (x \wedge x_i)$. In particular for the linear spline ($d = 1$) we have

$$K_1(x_j, x_i) = 1 + x_j x_i + \frac{1}{2} |x_j - x_i| (x_j \wedge x_i)^2 + \frac{(x_j \wedge x_i)^3}{3}$$

Again the kernel for n -dimensional splines with an infinite number of knots is the product of the n kernels for one-dimensional splines.

On the basis of this kernel, one can construct a spline approximation (using the techniques described in previous section) that has the form

$$f(x, \beta) = \sum_{i=1}^{\ell} \beta_i K(x, x_i)$$

11.6.3 B_d -Spline Approximations

In computational mathematics an important role belongs to the so-called B_d -spline approximations. There are two ways to define B_n splines: By iterative procedure or as a linear combination of regular splines.

Definition 1. Let us call the following function B_0 spline (B spline of order 0):

$$B_0(u) = \begin{cases} 1 & \text{if } |u| \leq 0.5, \\ 0 & \text{if } |u| > 0.5. \end{cases}$$

The B_d spline of order d we define as a convolution of two functions: B_{d-1} spline and B_0 spline:

$$B_d(u) = \int_{-\infty}^{\infty} B_{d-1}(u-t)B_0(t) dt. \quad (11.50)$$

Definition 2. The $B_d(u)$ spline has the following construction:

$$B_d(u) = \sum_{r=0}^{d+1} \frac{(-1)^r}{r!} C_{d+1}^r \left(u + \frac{d+1}{2} - r \right)_+^d.$$

One can show that both definitions describe the same object.

Using B_d splines, one can approximate functions by expansion:

$$f(x, \beta) = \sum_{i=1}^N \beta_i B_d(x - t_i),$$

where t_i , $i = 1, \dots, N$, defines knots of expansion. Since this expansion has the form of an inner product, the kernel that generates B-spline expansion is

$$K(x, x_i) = \sum_{k=1}^N B_d(x - t_k) B_d(x_i - t_k).$$

There is a good approximation for a B_d spline:

$$B_d(u) \approx \sqrt{\frac{6}{\pi(d+1)}} \exp \left\{ -\frac{6u^2}{d+1} \right\}. \quad (11.51)$$

The approximation becomes better with increasing d, but is surprisingly good even for $d = 1$. See Fig. 11.3.

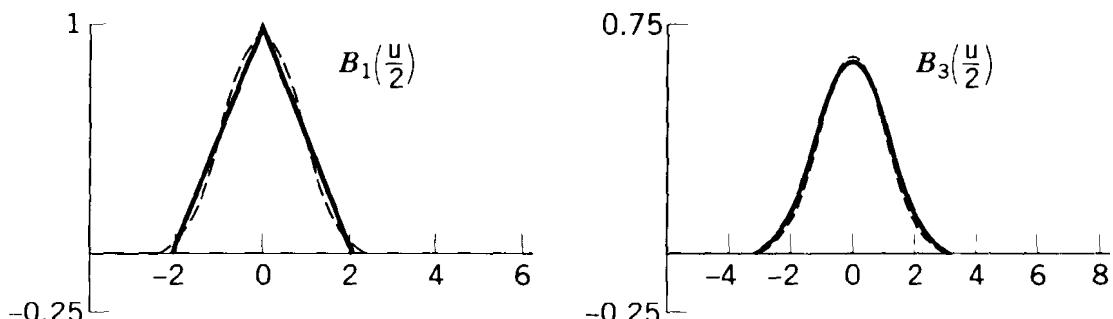


FIGURE 11.3. B_d -splines and their approximations by the Gaussians.

11.6.4 B_d Splines with an Infinite Number of Knots

Consider now an expansion of the form

$$f(x) = \int_{-\infty}^{\infty} \phi(t) B_d(x - t) dt, \quad (11.52)$$

where $B_d(x - t)$ is a B_d spline and $\phi(t)$ is a function that defines the approximation $f(x)$. For any fixed value x the expression (11.52) describes the inner product between two functions. Therefore the inner product between two B_d splines that defines the desired kernel has the form

$$\begin{aligned} K(x_i, x_j) &= \int_{-\infty}^{\infty} B_d(x_i - t) B_d(x_j - t) dt \\ &= \int_{-\infty}^{\infty} B_d(x_i - t) B_d(t - x_j) dt \\ &= B_{2d+1}(x_i - x_j) \end{aligned}$$

(the second equality is due to the fact that B_d splines are symmetric functions, and the third equality is due to definition 1 of the $B_d(x)$ splines). Thus the kernels for constructing one-dimensional B_d splines are defined by a B_{2d+1} spline.

Again, the kernel for n -dimensional B_d splines is the product of n one-dimensional kernels:

$$K(x_i, x_j) = \prod_{r=1}^n B_{2d+1}(x_i^r - x_j^r).$$

Taking into account approximation (11.51), we obtain that

$$K(x_i, x_j) = \prod_{r=1}^n B_{2d+1}(x_i^r - x_j^r) \approx \left(\frac{3}{\pi(d+1)} \right)^{n/2} \exp \left\{ - \frac{3|x_i - x_j|^2}{d+1} \right\}$$

Thus, the kernel for constructing B_d splines can be approximated by Gaussian function.

11.7 KERNELS GENERATING FOURIER EXPANSIONS

An important role in signal processing belongs to Fourier expansions. In this section we construct kernels for SV Fourier expansions in multidimensional spaces. As before we start with the one-dimensional case.

Suppose we would like to analyze a one-dimensional signal in terms of Fourier series expansion.

Let us map the input variable x into the $(2N+1)$ -dimensional vector

$$u = (1/\sqrt{2}, \sin x, \dots, \sin Nx, \cos x, \dots, \cos Nx).$$

Then for any fixed x the Fourier expansion can be considered as the inner product in this $(2N+1)$ -dimensional feature space:

$$f(x) = (a * u) = \frac{a_0}{\sqrt{2}} + \sum_{k=1}^N (a_k \sin kx + b_k \cos kx). \quad (11.53)$$

Therefore the inner product of two vectors in this space has the form

$$K_N(x, x_i) = \frac{1}{2} + \sum_{k=1}^N (\sin kx \sin kx_i + \cos kx \cos kx_i).$$

After obvious transformations and taking into account Dirichlet function (see Chapter 6, Section 6.5), we obtain

$$K_N(x, x_i) = \frac{1}{2} + \sum_{k=1}^N \cos k(x - x_i) = \frac{\sin \frac{(2N+1)}{2}(x - x_i)}{\sin \frac{(x - x_i)}{2}}$$

To define the signal in terms of the Fourier expansion, the SV machine uses the representation

$$f(x, \beta) = \sum_{i=1}^{\ell} \beta_i K_N(x, x_i).$$

Again, to construct the SV machine for the d -dimensional vector space $x = (x^1, \dots, x^n)$, it is sufficient to use the generating kernel that is a product of one-dimensional kernels:

$$K(x, x_i) = \prod_{k=1}^n K(x^k, x_i^k).$$

11.7.1 Kernels for Regularized Fourier Expansions

In Section 6.5, when we considered approximation of the functions by Fourier expansions, we pointed out that the Dirichlet kernel does not have good approximation properties. Therefore we considered two other (regularized) ker-

nels: the Fejer kernel and the Jackson kernel. The following introduces two new kernels that we use for approximation of the multidimensional functions with SV machines.

Consider the following regularized Fourier expansion:

$$f(x) = \frac{a_0}{\sqrt{2}} + \sum_{k=1}^{\infty} q^k (a_k \cos kx + b_k \sin kx), \quad 0 < q < 1,$$

where a_k and b_k are coefficients of the Fourier expansion. This expansion differs from expansion (11.53) by multipliers q^k that provide a mode of regularization (see Fig. 11.4). The corresponding kernel for this regularized expansion is

$$\begin{aligned} K(x_i, x_j) &= \frac{1}{2} + \sum_{k=1}^{\infty} q^k (\cos kx_i \cos kx_j + \sin kx_i \sin kx_j) \\ &= \frac{1}{2} + \sum_{k=1}^{\infty} q^k \cos k(x_i - x_j) = \frac{1 - q^2}{2(1 - 2q \cos(x_i - x_j) + q^2)}. \end{aligned}$$

(For the last equality see Gradshteyn and Ryzhik (1980).)

Consider also the following regularization of the Fourier expansion:

$$f(x) = \frac{a_0}{\sqrt{2}} + \sum_{k=1}^{\infty} \frac{a_k \cos kx + b_k \sin kx}{1 + \gamma^2 k^2},$$

where a_k and b_k are coefficients of the Fourier expansion (see Fig. 11.5). This regularizer provides another mode of regularization than the first one. For

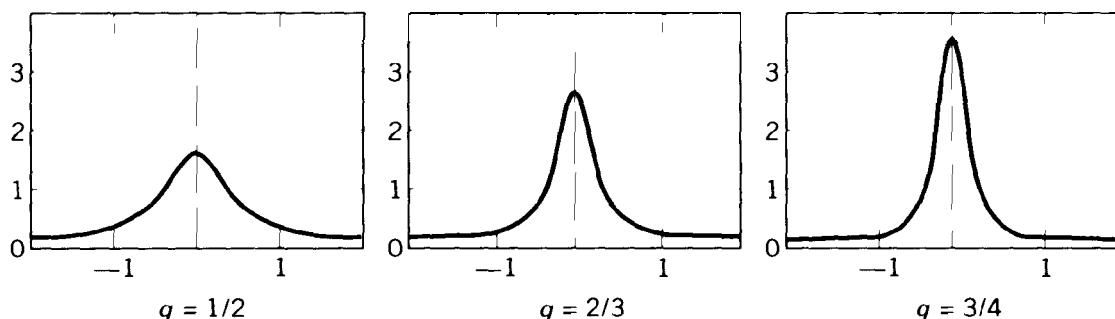
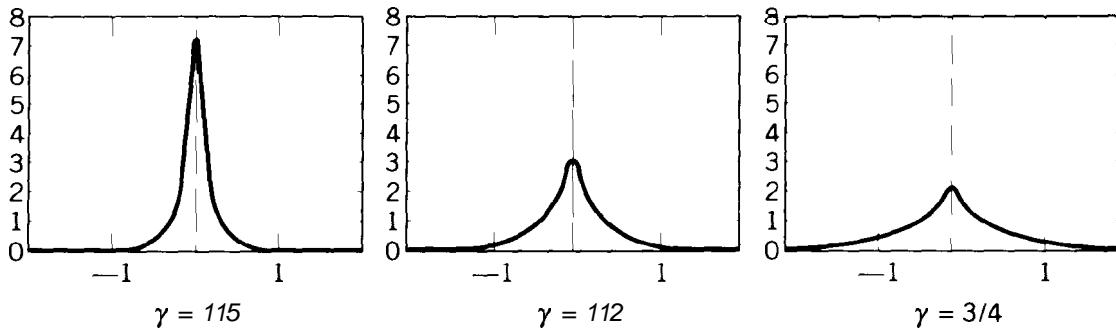


FIGURE 11.4. Kernels for various values of q .

FIGURE 11.5. Kernels for various values of γ .

this type of regularized Fourier expansion we have the following kernel:

$$\begin{aligned} K(x_i, x_j) &= \frac{1}{2} + \sum_{k=1}^{\infty} \frac{\cos kx_i \cos kx_j + \sin kx_i \sin kx_j}{1 + y^2 k^2} \\ &= \frac{\pi}{2\gamma} \frac{\operatorname{ch} \frac{\pi - |x_i - x_j|}{\gamma}}{\operatorname{sh} \frac{\pi}{\gamma}}, \quad 0 \leq |x_i - x_j| \leq 2\pi. \end{aligned}$$

(For last equality see Gradshteyn and Ryzhik (1980).)

Again the kernels for multidimensional Fourier expansion is the product of the kernels for one-dimensional Fourier expansions.

11.8 THE SUPPORT VECTOR ANOVA DECOMPOSITION (SVAD) FOR FUNCTION APPROXIMATION AND REGRESSION ESTIMATION

The kernels defined in previous sections can be used both for approximating multidimensional functions and for estimating multidimensional regressions. However, they can define a too rich set of functions. Therefore to control generalization, one needs to make a structure on this set of functions in order to choose the function from an appropriate element of the structure. Note also that when the dimensionality of the input space is large (say 100), the values of an n -dimensional kernel (which is the product of an n one-dimensional kernels) can have an order of magnitude q^n . These values are inappropriate for both cases when $q > 1$ and $q < 1$.

Classical statistics considered the following structure on the set of multidimensional functions from L_2 , the so-called **ANOVA** (acronym for analysis of variances) decomposition.

Suppose that an n -dimensional function $f(\mathbf{x}) = f(x^1, \dots, x^n)$ is defined on the set $I \times I \times \dots \times I$, where I is a finite or infinite interval.

The ANOVA decomposition of function $f(\mathbf{x})$ is an expansion

$$f(x^1, \dots, x^n) = F_0 + F_1(x^1, \dots, x^n) + F_2(x^1, \dots, x^n) + \dots + F_n(x^1, \dots, x^n),$$

where

$$\begin{aligned}
 F_0 &= C, \\
 F_1(x^1, \dots, x^n) &= \sum_{1 \leq k \leq n} \psi_k(x^k), \\
 F_2(x^1, \dots, x^n) &= \sum_{1 \leq k_1 < k_2 \leq n} \phi_{k_1, k_2}(x^{k_1}, x^{k_2}), \\
 &\dots \\
 F_r(x^1, \dots, x^n) &= \sum_{1 \leq k_1 < k_2 < \dots < k_r \leq n} \phi_{k_1, \dots, k_r}(x^{k_1}, x^{k_2}, \dots, x^{k_r}), \\
 F_n(x^1, \dots, x^n) &= \phi_{k_1, \dots, k_n}(x^1, \dots, x^n).
 \end{aligned}$$

The classical approach to **ANOVA** decompositions has a problem with the exponential explosion of the number of summands with increasing order of approximation. In Support Vector techniques, we do not have this problem. To construct the kernel for the **ANOVA** decomposition of order p using a sum of products of one-dimensional kernels $K(x^i, x_r^i)$, $i = 1, \dots, n$

$$K_p(x, x_r) = \sum_{1 \leq i_1 < i_2 < \dots < i_p \leq n} K(x^{i_1}, x_r^{i_1}) \times \dots \times K(x^{i_p}, x_r^{i_p})$$

one can introduce a recurrent procedure for computing $K_p(x, x_r)$, $p = 1, \dots, n$.

Let us denote

$$K^s(x, x_r) = \sum_{i=1}^n K^s(x^i, x_r^i).$$

One can easily check that the following recurrent procedure define the kernels $K_p(x, x_r)$, $p = 1, \dots, n$:

$$\begin{aligned}
 K_0(x, x_r) &= 1, \\
 K_1(x, x_r) &= \sum_{1 \leq i \leq n} K(x^i, x_r^i) = K^1(x, x_r), \\
 K_2(x, x_r) &= \sum_{1 \leq i_1 < i_2 \leq n} K(x^{i_1}, x_r^{i_1}) K(x^{i_2}, x_r^{i_2}) \\
 &= \frac{1}{2} [K_1(x, x_r) K^1(x, x_r) - K^2(x, x_r)], \\
 K_3(x, x_r) &= \sum_{1 \leq i_1 < i_2 < i_3 \leq n} K(x^{i_1}, x_r^{i_1}) K(x^{i_2}, x_r^{i_2}) K(x^{i_3}, x_r^{i_3}) \\
 &= \frac{1}{3} [K_2(x, x_r) K^1(x, x_r) - K_1(x, x_r) K^2(x, x_r) + K^3(x, x_r)].
 \end{aligned}$$

In general case we have[†]

$$K_p(x, x_r) = \frac{1}{p} \sum_{s=1}^p (-1)^{s+1} K_{p-s}(x, x_r) K^s(x, x_r).$$

To construct **SV ANOVA** decomposition with orthogonal expansion, one has to use one-dimensional generating kernels constructed from an orthogonal basis (e.g., the kernel defined by Eq. (11.42) if one considers an infinite interval \mathbf{I} or corresponding kernels for regularized Fourier expansion if one considers a finite interval \mathbf{I}).

Using such kernels, and the **SV** method with L_2 loss function, one can obtain an approximation of any order.

However, it is important to perform **ANOVA** decomposition for approximations that is based on RBF or splines with infinite number of knots. For such approximations the **ANOVA** decomposition is not orthogonal and one can approximate the target function well using only one term $F_p(x^1, \dots, x^n)$ (of appropriate order). Using the **SV** method with L_1 ϵ -insensitive loss-function and the corresponding generating kernel $K_p(x, x_i)$ one obtains such approximations.

11.9 SV METHOD FOR SOLVING LINEAR OPERATOR EQUATIONS

This section uses the **SV** method for solving linear operator equations

$$Af(t) = F(x), \quad (11.54)$$

where operator A realizes a one-to-one mapping from a Hilbert space E_1 into a Hilbert space E_2 .

We will solve equations in the situation where instead of function $F(x)$ on the right-hand side of (11.54) we are given measurements of this function (generally with errors)

$$(x_1, F_1), \dots, (x_\ell, F_\ell). \quad (11.55)$$

It is necessary to estimate the solution of Eq. (11.54) from the data (11.55).

The following shows that the **SV** technique realizes the classical ideas of solving ill-posed problems where the choice of the kernel is equivalent to the choice of the regularization functional. Using this technique, one can solve operator equations in high-dimensional spaces.

11.9.1 The SV Method

In Appendix to Chapter 1, we formulated the regularization method of solving operator equations, where in order to solve operator Eq. (11.54) one

[†]"A New Method for Constructing Artificial Neural Networks" Technical Report ONR Contract N00014-94-C-0186 Data Item A002. May 1, 1995. Prepared by C. Burges and V. Vapnik.

minimizes the functional

$$R_\gamma(f, F) = \rho^2(Af, F) + \gamma W(f),$$

where solution belongs to some compact $W(f) \leq C$ (C is unknown constant). When one solves operator Eq. (11.54) using data (11.55) one considers the functional

$$R_\gamma(f, F) = \frac{1}{\ell} \sum_{i=1}^{\ell} L(Af(t)|_{x_i} - F_i) + \gamma(Pf * Pf)$$

with some loss function $L(Af - F)$ and regularizer of the form

$$W(f) = (Pf * Pf)$$

defined by some nongenerating operator P . Let

$$\begin{aligned} \varphi_1(t), \dots, \varphi_n(t), \dots \\ \lambda_1, \dots, \lambda_n, \dots \end{aligned}$$

be eigenfunctions and eigenvalues of the selfconjugate operator $P * P$

$$P^* P \varphi_i = \lambda_i \varphi_i.$$

Consider the solution of Eq. (11.54) as the expansion

$$f(t) = \sum_{k=1}^{\infty} \frac{w_k}{\sqrt{\lambda_k}} \varphi_k(t).$$

Putting this expansion into functional $R_\gamma(f, F)$ we obtain

$$R_\gamma(f, F) = \frac{1}{\ell} \sum_{i=1}^{\ell} L \left(A \left\{ \sum_{k=1}^{\infty} \frac{w_k}{\sqrt{\lambda_k}} \varphi_k(t) \right\} \Big|_{x_i} - F_i \right) + \gamma \sum_{k=1}^{\infty} w_k^2.$$

Denoting

$$\phi_k(t) = \frac{\varphi_k(t)}{\sqrt{\lambda_k}}$$

we can rewrite our problem in the familiar form: minimize the functional

$$R_\gamma(w, F) = \frac{1}{\ell} \sum_{i=1}^{\ell} L(|A(w * \Phi(t))|_{x=x_i} - F_i) + \gamma(w * w)$$

in the set of functions

$$f(t, w) = \sum_{r=1}^{\infty} w_r \phi_r(t) = (w * \Phi(t)), \quad (11.56)$$

where we denote

$$\begin{aligned} \mathbf{w} &= (w_1, \dots, w_N, \dots), \\ \Phi(t) &= (\phi_1(t), \dots, \phi_N(t), \dots). \end{aligned} \quad (11.57)$$

The operator A maps this set of functions into the set of functions

$$F(x, \mathbf{w}) = Af(t, \mathbf{w}) = \sum_{r=1}^{\infty} w_r A \phi_r(t) = \sum_{r=1}^{\infty} w_r \psi_r(x) = (\mathbf{w} * \Psi(x)), \quad (11.58)$$

which is linear in the feature space

$$\Psi(x) = (\psi_1(x), \dots, \psi_N(x), \dots),$$

where

$$\psi_r(x) = A \phi_r(t).$$

To find the solution of Eq. (11.54) in a set of functions $f(t, \mathbf{w})$ (to find the vector coefficients \mathbf{w}), one can minimize the functional

$$D(F) = C \sum_{i=1}^{\ell} (|F(x_i, \mathbf{w}) - F_i|_\varepsilon)^k + (\mathbf{w} * \mathbf{w}), \quad k = 1, 2$$

in the image space that is in the space of functions $F(x, \mathbf{w})$.

Let us define the generating kernel in the image space

$$K(x_i, x_j) = \sum_{r=0}^{\infty} \psi_r(x_i) \psi_r(x_j) \quad (11.59)$$

and the so-called cross-kernel function

$$\mathcal{K}(x_i, t) = \sum_{r=0}^{\infty} \psi_r(x_i) \phi_r(t) \quad (11.60)$$

(here we suppose that the operator A is such that the right-hand side converges uniformly for x and t).

Note that in this case the problem of finding the solution to the operator equation (finding the corresponding vector of coefficients \mathbf{w}) is equivalent to the problem of finding vector \mathbf{w} for the linear regression function (11.58) in the image space using measurements (11.55).

Let us solve this regression problem using the quadratic optimization **SV** technique. That is, using kernel (11.59), one can find both the support vectors x_i , $i = 1, \dots, N$, and the corresponding coefficients $\alpha_i^* - \alpha_i$ that define the vector \mathbf{w} for the **SV** regression approximation:

$$\mathbf{w} = \sum_{i=1}^N (\alpha_i^* - \alpha_i) \Psi(x_i).$$

Note that the same coefficients w along with regression in image space define the approximation to the desired solution in preimage space. Therefore putting these coefficients in the expression (11.%), one obtains

$$f(t, \mathbf{a}, \mathbf{a}^*) = \sum_{i=1}^N (\alpha_i^* - \alpha_i) \mathcal{K}(x_i, t).$$

That is, we find the solution to our problem of solving the operator equation using the cross-kernel function as an expansion on support vectors.

Therefore in the **SV** technique for solving operator equations the choice of the kernel function is equivalent to the choice of the regularization functional. The cross-kernel function is constructed taking into account the regularization functional and the operator.

Therefore in order to solve the linear operator equation using the **SV** method:

1. Define the corresponding regression problem in image space.
2. Construct the kernel function $K(x_i, x_j)$ for solving the regression problem using the **SV** method.
3. Construct the cross-kernel function $\mathcal{K}(x_i, t)$.
4. Using the kernel function $K(x_i, x_j)$, solve the regression problem by the **SV** method' (i.e., find the support vectors x_i^* , $i = 1, \dots, N$, and the corresponding coefficients $\beta_i = (\alpha_i^* - \alpha_i)$, $i = 1, \dots, N$).
5. Using these support vectors and the corresponding coefficients, define the solution

$$f(t) = \sum_{r=1}^N \beta_r \mathcal{K}(x_r, t). \quad (11.61)$$

Steps 1–3 (constructing regression problem, constructing kernel in image space, and constructing corresponding cross-kernel function) reflect the specific problem at hand (they depend on operator A). Steps 4 and 5 (solving the regression problem by **SV** machine and constructing the solution to the desired problem) are routine.

The main problem with solving operator equations using the **SV** technique is for a given operator equation to obtain both the explicit expression for the kernel function in image space and the explicit expression for the corresponding cross-kernel function. In the next section, which is devoted to solving special integral equations that form the (multidimensional) density estimation problem, we construct such pairs.

In Chapter 13, which is devoted to the application of the **SV** method to real-function estimation problems, we construct such a pair for another

[†]Note that since in (11.58) coefficient $b = 0$, the constraint $\sum \alpha_i^* = \sum \alpha_i$ in the optimization problem should be removed.

problem of solving operator equations—for solving the Radon equation for positron emission tomography (PET).

For solving these operator equations, we will use functions of Hilbert spaces that are defined in a form slightly different from the form considered above; that is, we will look for the solution of the operator equation

$$Af(t) = F(x)$$

in the set of functions

$$f(t) = \int g(\tau) \psi(t, \tau) d\tau,$$

where $\psi(t, \tau)$ is a given function and $g(\tau)$ can be any function from some Hilbert space. To find the solution means to estimate the function $g(\tau)$ (instead of infinite dimensional vector as above). Let us denote

$$A\psi(t, \tau) = \phi(x, \tau)$$

and rewrite our equation as follows:

$$\int g(\tau) \phi(x, \tau) d\tau = F_g(x).$$

(assume that $\phi(x, \tau)$ is such that for any fixed x function, $\phi(x, \tau)$ belongs to L_2). Since for any fixed x the left-hand side of the equation is the inner product between two functions of the Hilbert space, one can construct the kernel function

$$K(x_i, x_j) = \int \phi(x_i, \tau) \phi(x_j, \tau) d\tau$$

and cross-kernel function

$$\mathcal{K}(x, t) = \int \psi(t, \tau) \phi(x, \tau) d\tau.$$

These functions are used to obtain the solution:

$$f(t) = \sum_j \beta_j \mathcal{K}(x_j, t),$$

where coefficients $\beta_j = \alpha_j^* - \alpha_j$ are found using standard SV techniques with the kernel $K(x_i, x_j)$.

This solution of the operator equation reflects the following regularization idea: It minimizes the functional

$$R(g) = C \sum_{i=1}^{\ell} |F_g(x_i) - F_i|_{\epsilon}^k + (g * g), \quad k = 1, 2.$$

In the remaining part of this section we discuss some additional opportunities of the SV technique that come from the ability to control the ϵ -insensitivity.

11.9.2 Regularization by Choosing Parameters of ε_i -Insensitivity

Until now, when we considered the problem of solving operator equations, we ignored the fact that it can be ill-posed—for example, if our equation is a Fredholm integral equation of the first kind (see Chapter 1, Section 1.11). Now this feature is the subject of our interest.

Chapter 7 considered methods of solving stochastic ill-posed problems by using the regularization method (see also Appendix to Chapter 1). According to the regularization method, in order to find a solution of the operator equation

$$Af = F \quad (11.62)$$

(that forms an ill-posed problem) in a situation where instead of the right-hand side of the equation the approximation F_ℓ is given, one has to minimize the functional

$$R(f) = \|Af - F_\ell\|^2 + \gamma_\ell W(f)$$

in a set of function $\{f\}$. In this functional the term $W(f)$ is the regularization functional and the parameter γ_ℓ is the regularization constant. One of the most important questions in solving an ill-posed problem is how to choose the value γ_ℓ .

To choose this constant Morozov (1984) suggested the so-called **residual principle**: Suppose that one knows that the accuracy of the approximating function F_ℓ obtained from the data does not exceed ε ; then one has to minimize the regularization functional $W(f)$ subject to constraint

$$\|Af - F_\ell\| \leq \varepsilon. \quad (11.63)$$

By using the ε -insensitive loss function the SV method of solving operator equation realizes this idea in the stronger form: For sufficiently large C it minimizes the regularization functional (norm of the vector of coefficients of linear function in feature space) subject to constraint

$$|F(x_i) - F_i| \leq \varepsilon_i, \quad i = 1, \dots, \ell$$

Such a mode of regularization is used when one has information on the accuracy ε_i of the measurements in any point of approximation. As we will see in the next section, in the problem of density estimation as well as in the PET problem discussed in Chapter 13 simultaneously with data describing the right-hand side of the equation, one can estimate the accuracy of obtained data in any specific point. In other words, one has to solve the operator equation given the triples

$$(x_1, F_1, \varepsilon_1), \dots, (x_\ell, F_\ell, \varepsilon_\ell).$$

Using various values for ε -insensitivity for various points (vectors) x , one can control the regularization processes better.

11.10 SV METHOD OF DENSITY ESTIMATION

Let us apply the SV method for solving linear operator equations to the problem of density estimation. First we obtain a method for estimating one-dimensional densities, and then using the standard approach we generalize this method for estimating multidimensional densities. In this subsection, in order to simplify the notations we consider the problem of density estimation on the interval $[0,1]$.

As was shown in Chapter 1, Section 1.8, the problem of density estimation is a problem of solving the integral equation

$$\int_0^1 \theta(x-t)p(t)dt = F(x), \quad (11.64)$$

where instead of distribution function $F(x)$ the i.i.d. data are given:

$$x_1, \dots, x_\ell.$$

Using these data, one constructs the empirical distribution function[†]

$$F_\ell(x) = \frac{1}{\ell} \sum_{i=1}^{\ell} \theta(x - x_i)$$

and instead of the right-hand side of (11.64) considers the measurements

$$(x_1, F_\ell(x_1)), \dots, (x_\ell, F_\ell(x_\ell)). \quad (11.65)$$

One also adds the boundary conditions

$$(0, 0), (1, 1).$$

It is easy to check that for any point x^* the random value $F_\ell(x^*)$ is unbiased and has the standard deviation

$$\sigma^* = \sqrt{\frac{1}{\ell} F(x^*)(1 - F(x^*))} \leq \frac{1}{2\sqrt{\ell}}.$$

Let us characterize the accuracy of approximation of the value $F(x_i)$ by the value $F_\ell(x_i)$ with

$$\varepsilon_i^* = c\sigma_i = c\sqrt{\frac{1}{\ell} F(x_i)(1 - F(x_i))},$$

where c is some constant.

[†]Empirical distribution function

$$F_\ell(x) = \frac{1}{\ell} \sum_{i=1}^{\ell} \theta(x^1 - x_i^1) \dots \theta(x^n - x_i^n)$$

in multidimensional case $\mathbf{x} = (x^1, \dots, x^n)$.

Since the distribution function $F(x)$ is unknown, let us approximate ε_i^* by the value

$$\varepsilon_i = \sqrt{\frac{1}{\ell}(F_\ell(x_i) + \delta)(1 - F_\ell(x_i) + \delta)},$$

where $\delta > 0$ is some small parameter. Therefore one constructs triplets:

$$(x_1, F_\ell(x_1), \varepsilon_1), \dots, (x_\ell, F_\ell(x_\ell), \varepsilon_\ell). \quad (11.66)$$

11.10.1 Spline Approximation of a Density

We are looking for the solution of Eq. (11.64) as an expansion in a spline function with an infinite number of nodes.

That is we approximate the unknown density by the function

$$p(t) = \int_0^1 g(\tau)(t - \tau)_+^d d\tau + \sum_{k=0}^d a_k t^k,$$

where $g(\tau)$ is a function to be estimated and a_k , $k = 0, 1, \dots, d$, are the parameters to be estimated. To simplify formulas below we consider linear splines ($d = 1$). The case of $d \neq 1$ is completely analogous.

According to the SV method described in the previous section, to solve linear operator equations we have to perform five steps, among which the first three steps

1. Define the corresponding regression problem in image space
2. Construct the kernel function $K(x, x_i)$
3. Construct the cross-kernel function $\mathcal{K}(x, t)$

are specific for the problem, while the last two steps are routine.

Below we consider the first three steps of solving the density estimation problem.

Step 1. We define the regression problem as a problem of approximation of the following function $F(x)$ in image space:

$$\begin{aligned} F(x) &= \int_0^1 g(\tau) \left[\int_0^x (t - \tau)_+ dt \right] d\tau + \int_0^x (a_1 t + a_0) dt \\ &= \int_0^1 g(\tau) \left[\frac{(x - \tau)_+^2}{2} \right] d\tau + \frac{a_1 x^2}{2} + a_0 x \end{aligned}$$

using the data (11.66).

Step 2. Since the last formula can be considered as an inner product, we

construct the following kernel in image space:

$$\begin{aligned}
 K(x_i, x_j) &= \frac{1}{4} \int_0^1 (x_i - \tau)_+^2 (x_j - \tau)_+^2 d\tau + \frac{x_i^2 x_j^2}{4} + x_i x_j \\
 &= \frac{1}{4} \int_0^{(x_i \wedge x_j)} (x_i - \tau)^2 (x_j - \tau)^2 d\tau + \frac{x_i^2 x_j^2}{4} + x_i x_j \\
 &= |x_i - x_j|^2 \frac{(x_i \wedge x_j)^3}{12} + |x_i - x_j| \frac{(x_i \wedge x_j)^4}{8} + \frac{(x_i \wedge x_j)^5}{20} + \frac{x_i^2 x_j^2}{4} + x_i x_j,
 \end{aligned} \tag{11.67}$$

where we denoted by $(x_i \wedge x_j)$ the minimum of two values x_i and x_j .

Step 3. We evaluate the cross-kernel function

$$\begin{aligned}
 \mathcal{K}(x_i, t) &= \frac{1}{2} \int_0^1 (x_i - \tau)_+^2 (t - \tau)_+ d\tau + \frac{x_i^2 t}{2} + x_i \\
 &= \frac{1}{2} \int_0^{(x_i \wedge t)} (x_i - \tau)^2 (t - \tau) d\tau + \frac{x_i^2 t}{2} + x_i = \frac{x_i^2 t}{2} + x_i \\
 &\quad + \frac{x_i^2 t (x_i \wedge t)}{2} - (2x_i t + x_i^2) \frac{(x_i \wedge t)^2}{4} + (2x_i + t) \frac{(x_i \wedge t)^3}{6} - \frac{(x_i \wedge t)^4}{8}.
 \end{aligned} \tag{11.68}$$

Using kernel (11.67) and the triplets (11.66), we obtain the support vectors x_k , $k = 1, \dots, N$, and the corresponding coefficients $\beta_k^0 = \alpha_k^* - \alpha_k$, $k = 1, \dots, N$, that define the SV regression approximation:

$$F(x) = \sum_{k=1}^N \beta_k^0 K(x_k, x).$$

These parameters and cross-kernel function (11.68) define the desired SV approximation of the density

$$p(t) = \sum_{k=1}^N \beta_k^0 \mathcal{K}(x_k, t).$$

To solve the multidimensional problem of density estimation, one has to construct a multidimensional kernel function and a multidimensional cross-kernel function, which are products of one-dimensional kernel functions and one-dimensional cross-kernel functions.

11.10.2 Approximation of a Density with Gaussian Mixture

Consider the same method of density estimation in a set of functions defined in $[0, \infty)$ as Gaussian mixtures:

$$p(t) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} g(\tau) \exp \left\{ -\frac{(t - \tau)^2}{2\sigma^2} \right\} d\tau, \tag{11.69}$$

where the functions $g(\tau)$ defining the approximations belong to L_2 , σ is a fixed parameter, and values (vectors) t are nonnegative.

Let us start with the one-dimensional case. Consider the regression space for our density estimation problem

$$F_g(x) = \int_{-\infty}^{\infty} g(\tau) \left(\frac{1}{\sqrt{2\pi}\sigma} \int_0^x \exp \left\{ -\frac{(t-\tau)^2}{2\sigma^2} \right\} dt \right) d\tau$$

Since for any fixed x this function has a structure of the inner product between two functions in Hilbert space, one can define the kernel function

$$K(x_i, x_j) = \frac{1}{2\pi\sigma^2} \int_{-\infty}^{\infty} d\tau \int_0^{x_i} \exp \left\{ -\frac{(t_1-\tau)^2}{2\sigma^2} \right\} dt_1 \int_0^{x_j} \exp \left\{ -\frac{(t_2-\tau)^2}{2\sigma^2} \right\} dt_2, \quad (11.70)$$

and the cross-kernel function

$$\mathcal{K}(x_i, t) = \frac{1}{2\pi\sigma^2} \int_{-\infty}^{\infty} \left(\exp \left\{ -\frac{(x-\tau)^2}{2\sigma^2} \right\} \int_0^{x_i} \exp \left\{ -\frac{(t-\tau)^2}{2\sigma^2} \right\} dt \right) d\tau. \quad (11.71)$$

The important feature of the Gaussian mixture solution is that both the kernel function and the cross-kernel function have a simple expression in terms of erf functions

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$$

and the integral from erf functions

$$\text{interf}(x) = \int_0^x \text{erf}(x') dx'.$$

The erf function is a smooth function that tabulated on computers. One can also easily tabulate the integral of the erf function (let us call this function the interf function).

Let us compute the kernel function (11.70). By changing the order of integration, one obtains

$$\begin{aligned} K(x_i, x_j) &= \frac{1}{2\pi\sigma^2} \int_0^{x_i} \int_0^{x_j} dt_1 dt_2 \int_{-\infty}^{\infty} \exp \left\{ -\frac{(t_1-\tau)^2}{2\sigma^2} \right\} \exp \left\{ -\frac{(t_2-\tau)^2}{2\sigma^2} \right\} d\tau \\ &= \frac{1}{2\pi\sigma^2} \int_0^{x_i} \int_0^{x_j} dt_1 dt_2 \int_{-\infty}^{\infty} \exp \left\{ -\frac{(t_1-t_2)^2}{4\sigma^2} - \frac{\left(\frac{t_1+t_2}{2} - \tau \right)^2}{2\sigma^2} \right\} d\tau \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_0^{x_i} \int_0^{x_j} \exp \left\{ -\frac{(t_1-t_2)^2}{4\sigma^2} \right\} dt_1 dt_2 \end{aligned}$$

$$\begin{aligned}
&= \sqrt{2}\sigma \left(\int_0^{\frac{x_i}{2\sigma}} \operatorname{erf}(u) du + \int_0^{\frac{x_j}{2\sigma}} \operatorname{erf}(u) du - \int_0^{\frac{|x_i - x_j|}{2\sigma}} \operatorname{erf}(u) du \right) \\
&= \sqrt{2}\sigma \left[\operatorname{interf} \left(\frac{x_i}{2\sigma} \right) + \operatorname{interf} \left(\frac{x_j}{2\sigma} \right) - \operatorname{interf} \left(\frac{|x_i - x_j|}{2\sigma} \right) \right].
\end{aligned}$$

Analogously, one computes the cross-kernel function (11.71):

$$\begin{aligned}
\mathcal{K}(x_i, t) &= \frac{1}{2\pi\sigma^2} \int_0^{x_i} dx \int_{-\infty}^{\infty} \exp \left\{ -\frac{(x-t)^2}{4\sigma^2} - \frac{\left(\frac{x+t}{2} - \tau\right)^2}{2\sigma^2} \right\} d\tau \\
&= \frac{1}{\sqrt{2\pi}\sigma} \int_0^{x_i} \exp \left\{ -\frac{(x-t)^2}{4\sigma^2} \right\} dx \\
&= \frac{1}{\sqrt{2}} \left[\operatorname{erf} \left(\frac{x_i - t}{2\sigma} \right) + \operatorname{erf} \left(\frac{t}{2\sigma} \right) \right].
\end{aligned}$$

Using these kernel and cross-kernel functions in the general scheme for solving integral equations, one can estimate the density:

$$f(t) = \sum \beta_i \mathcal{K}(x_i, t), \quad (11.72)$$

where coefficients $\beta_i = \alpha_i^* - \alpha_i$ are obtained by solving the corresponding regression problem on the basis of the obtained kernel function.

The interesting feature of this solution (10.72) is that in spite of the fact that approximating functions are a mixture of Gaussians defined by (10.69), the basis functions $\mathcal{K}(x_i, t)$ in expansion (10.72) are not Gaussians. Figure 11.6 shows basis functions $\mathcal{K}(x_i, t)$ for $\sigma = 1$ and $x_i = 0.2, 0.4, 0.6, 0.8, 1$.

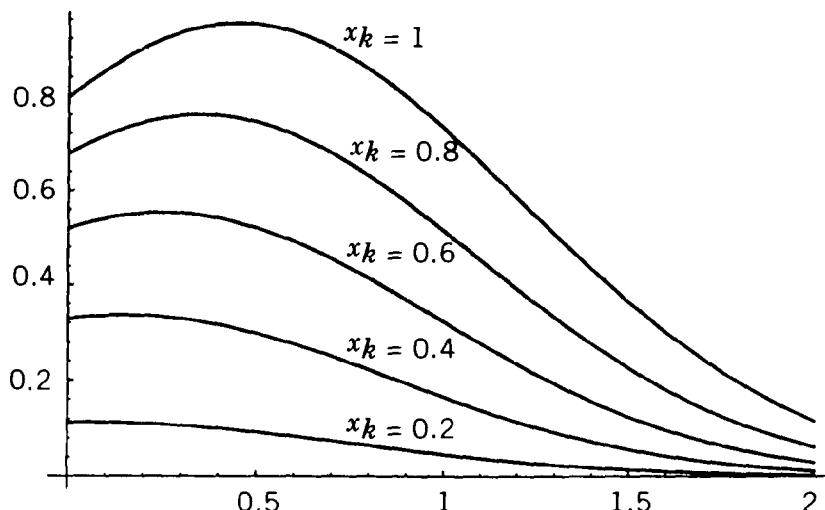


FIGURE 11.6. Cross-kernel function for a onedimensional density estimation in a mixture of Gaussians. For $\sigma = 1$, curves correspond to parameters $x_k = 0.2k$, $k = 1, \dots, 5$.

To estimate the multidimensional densities, one has to construct the multidimensional kernel function and the multidimensional cross-kernel function. As before, both multidimensional kernel function and multidimensional cross-kernel functions are products of corresponding one-dimensional functions.

It should be noted that this structure of multidimensional kernels is not necessarily valid for any operator equation. As we will see in Chapter 13, in particular, it is not valid in the case of the Radon tomography equation. To obtain the SV solution for the two-dimensional Radon equation, we will construct a two-dimensional kernel and a two-dimensional cross-kernel function.

11.11 ESTIMATION OF CONDITIONAL PROBABILITY AND CONDITIONAL DENSITY FUNCTIONS

11.11.1 Estimation of Conditional Probability Functions

In Chapter 7, Section 7.12 we considered the problem of estimating the conditional probability function using data

$$(y_1, z_1), \dots, (y_\ell, z_\ell), \quad y \in \{-1, 1\} \quad (11.73)$$

as a problem of solving the equation

$$\int_0^z p(y = 1|z) dF(z) = F(y = 1, z) \quad (11.74)$$

in the situation where the distribution functions $F(z)$, $F(y = 1, z)$ are unknown. To avoid the necessity of solving the high-dimensional integral equation (11.74) on the basis of data (11.73), we considered the method of estimating the conditional probability function along the line

$$z = z_0 + e(t - t_0)$$

passing through a point of interest z_0 , where the vector e defines the direction of the line. To estimate the conditional probability along this line, we split vectors z_i from (11.73) into two elements (t_i, u_i) , where $t_i = (z_i * e)$ is a projection of the vector z_i on the given direction e and u_i is an orthogonal complement of the vector et_i to the vector z_i . Let $z_0 = (t_0, u_0)$.

Therefore for the given direction e we constructed data

$$(y_1, t_1, u_1), \dots, (y_\ell, t_\ell, u_\ell), \quad y \in \{-1, 1\}, \quad (11.75)$$

which we used to solve the equation

$$\int_0^t p(y = 1|t, u_0) dF(t|u_0) = F(y = 1, t|u_0). \quad (11.76)$$

To solve this equation we introduced the approximations (Section 7.12)

$$F_\ell(t|u_0) = \sum_{i=1}^{\ell} \tau_i(u_0) \theta(t - t_i), \quad (11.77)$$

$$F_\ell(y = 1, t|u_0) \sum_{i=1}^{\ell} \tau_i(u_0) \theta(t - t_i) \delta(y_i), \quad (11.78)$$

$$\tau_i(u_0) = \frac{g_\gamma(\|u_i - u_0\|)}{\sum_{i=1}^{\ell} g_\gamma(\|u_i - u_0\|)}, \quad (11.79)$$

where $g_\gamma(u)$ is a Parzen kernel (with parameters of width γ) and $\delta(y_i) = 1$ if $y_i = 1$ and zero otherwise. In Chapter 7, Section 7.12 we described a method for solving this equation on the basis of approximations (11.77) and (11.78). However, we left undiscussed the problem of how to choose a good direction e .

Now let us discuss this problem. Our goal is to split the space into two subspaces: (1) a one-dimensional subspace that defines the most important direction for changing the conditional probability and (2) an orthogonal complement to this subspace. In our approximation we would like to take into account more accurately the important one-dimensional subspace.

To implement this idea we use the results of a solution to the pattern recognition problem to specify an important direction.[†]

First consider the case where a good decision rule is defined by a linear function. In this case it is reasonable to choose as an important direction one that is orthogonal to a separating hyperplane and as less important the directions that are parallel to a separating hyperplane. (See Fig. 11.7)

In general, the SV method solves a pattern recognition problem using a hyperplane in feature space, and therefore it is reasonable to choose the direction e defined by the vector that specifies the optimal hyperplane.

It is easy to check that if the inner product of two vectors in feature space Z is defined by the kernel $K(x_i, x_j)$ and α_i , $i = 1, \dots, \ell$, are coefficients that define the decision rule for a pattern recognition problem

$$f(x) = \theta \left\{ \sum_{i=1}^{\ell} y_i \alpha_i K(x_i, x) + b \right\},$$

then the quantities $t_i - t_0$ and $\|u_i - u_0\|$ can be defined using corresponding training data in input space

$$(y_1, x_1), \dots, (y_\ell, x_\ell),$$

[†]Note that the problem of pattern regression (regression estimation) is simpler than the problem of conditional probability (conditional density) estimation. Therefore, here we use the results of a solution to a simpler problem to solve a more difficult one.

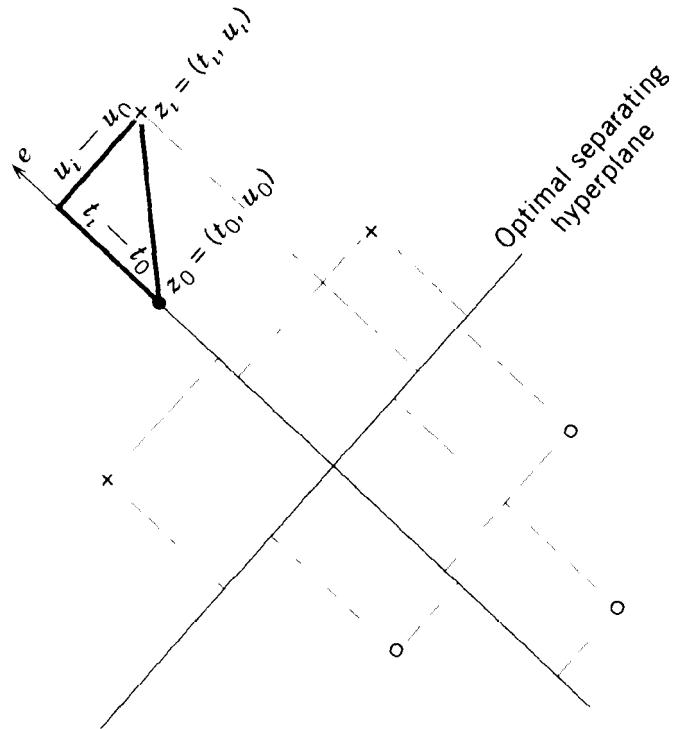


FIGURE 11.7. The line passing through the point of interest $z_0 = (t_0, u_0)$ in direction e defined by the optimal separating hyperplane.

as follows:

$$t_k - t_0 = \frac{\sum_{i=1}^{\ell} y_i \alpha_i [K(x_i, x_k) - K(x_i, x_0)]}{\sqrt{\sum_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j K(x_i, x_j)}} \quad (11.80)$$

$$\|u_k - u_0\| = \sqrt{K(x_k, x_k) + K(x_0, x_0) - 2K(x_k, x_0) - (t_k - t_0)^2}$$

Figure 11.8 demonstrates the result of estimating probability that the digit at the top of the figure is 3. In this picture the approximations $F_t(t|u_0)$, $F_t(y=3, t|u_0)$ and the obtained solution $p(y=3|t, u_0)$ are shown. The probability that the displayed digit is 3 is defined by the value of function $p(y=3|t, u_0)$ at the point $t = 0$. This probability is equal to 0.34 for example (a) and equal zero for example (b).

[†]Note that to estimate the conditional distribution functions (11.77), (11.78) one needs i.i.d. data (pairs t_i, u_i). If there are no additional training data such data can be obtained on the basis of the leave one out procedure. For the SV method it is sufficient to conduct this procedure only for support vectors.

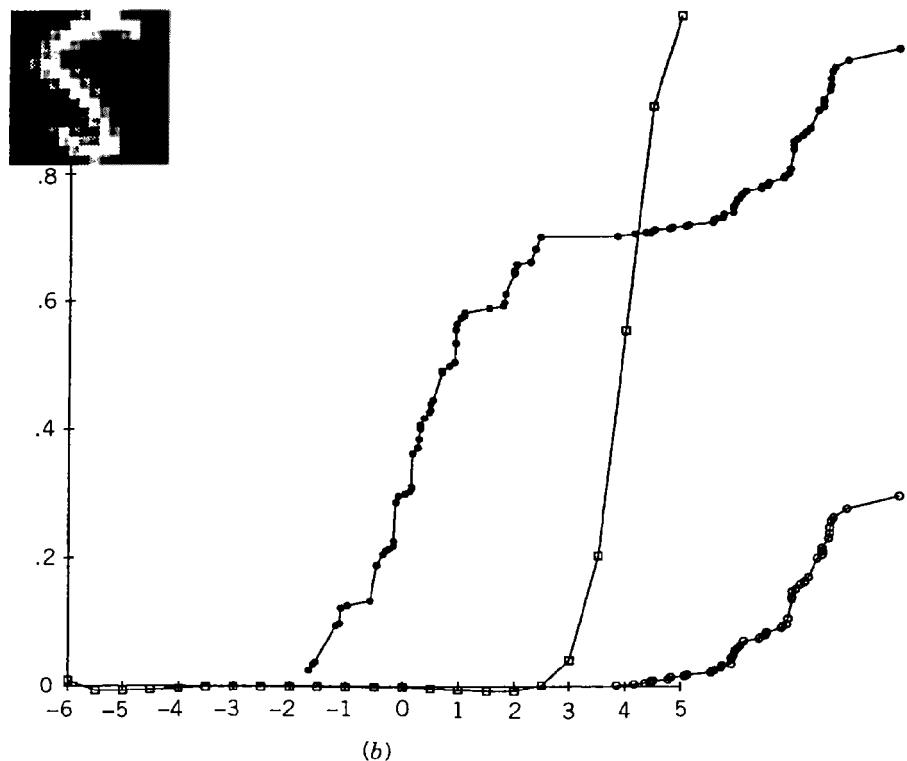
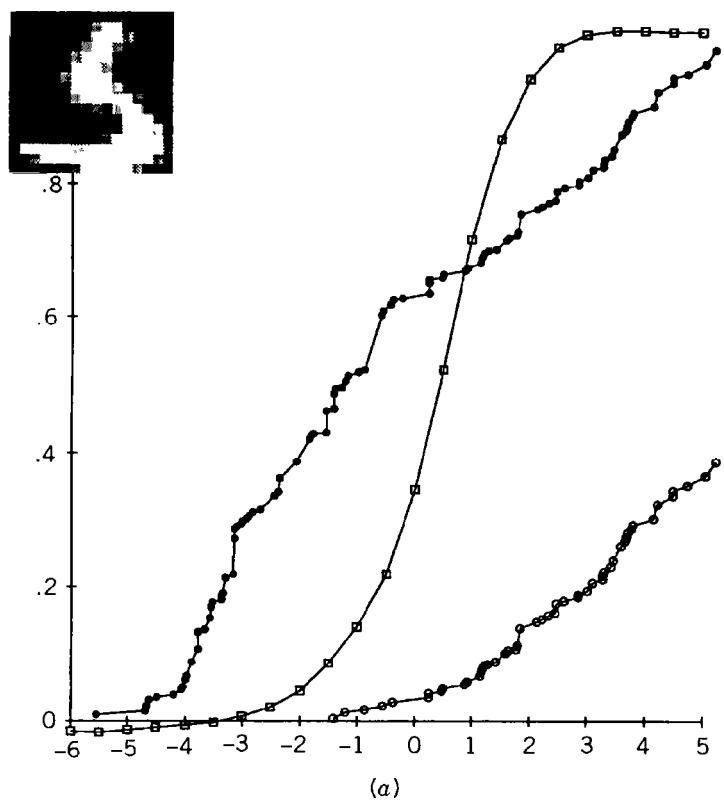


FIGURE 11.8. Approximations $F_\ell(t|u_0)$, $F_\ell(y=3, t|u_0)$ and the conditional probability along the line passing through point corresponding to the picture on the top of figure. The estimated probability that the corresponding digit is 3 equal to 0.34 for example (a) and zero for example (b).

11.11.2 Estimation of Conditional Density Functions

Section 7.11 considered the problem of estimating a conditional density function as a problem of solving the integral equation

$$\int_0^y \int_0^x p(y|x) dF(x) dy = F(y, x), \quad (11.81)$$

where y is real values. To solve this equation in a situation where the distribution functions $F(x)$ and $F(y, x)$ are unknown but the data

$$(y_1, x_1), \dots, (y_\ell, x_\ell) \quad (11.82)$$

are given, we used the same idea of estimating the desired function along the predefined line passing through a point of interest. Now we discuss how to define this line.

Suppose we have solved regression estimation problems using the SV technique. This means that we have mapped vectors x_i of our data (11.73) into a feature space

$$(y_1, z_1), \dots, (y_\ell, z_\ell)$$

where we construct the linear function

$$f(z) = (w * z).$$

In input space to this function there corresponds nonlinear regression

$$f(x) = \sum_{i=1}^{\ell} (\alpha_i^* - \alpha_i) K(x_i, x).$$

Consider as the important direction one that is orthogonal to linear regression function in feature space (i.e., one that is defined by vector of coefficients w of the estimated hyperplane). As above, vectors $z_i = z(x_i)$ are split into two elements, t_i which defines the position of the data on the important direction and u_i which is the orthogonal compliment to vector $z(x_i)$. Therefore we describe data $(y_1, z_1), \dots, (y_\ell, z_\ell)$ as follows:

$$(y_1, t_1, u_1), \dots, (y_\ell, t_\ell, u_\ell).$$

Let the vector $z_0 = (t_0, u_0)$ correspond to the point of our interest x_0 . Our goal is to use this data to obtain the solution of the equation

$$\int_0^y \int_0^t p(y|t, u_0) dF(t|u_0) dy = F(y, t|u_0). \quad (11.83)$$

Using expression (11.80) we construct the approximations

$$\begin{aligned} F_\ell(t|u_0) &= \sum_{i=1}^{\ell} \tau_i(u_0) \theta(t - t_i), \\ F_\ell(y, t|u_0) &= \sum_{i=1}^{\ell} \tau_i(u_0) \theta(t - t_i) \theta(y - y_i), \end{aligned} \quad (11.84)$$

where $\tau_i(u_0)$ is defined by (11.79).

In Section 7.11 we described a method for solving the Eq. (11.83) using approximations (11.84) that defines the conditional density along the line passing through a point of interest in a given direction.

We reduce the problem of solving our integral equation to the problem of solving a system of linear algebraic equations (7.87–7.89). Using this technique we obtain the desired approximation.[†]

11.12 CONNECTIONS BETWEEN THE SV METHOD AND SPARSE FUNCTION APPROXIMATION

In approximation theory (for example in wavelet approximation) the important problem is to approximate a given function $f(x)$ with a required accuracy using the smallest amount n of basis functions from a given collection of functions. In other words, it is required to construct the approximation of function

$$f(x) = \sum_{i=1}^n c_i \varphi_i(x) \quad (11.85)$$

satisfying the constraint

$$\|f(x) - \sum_{i=1}^n c_i \varphi_i(x)\|^2 \leq \varepsilon$$

using the smallest number of nonzero coefficients c_i .

Chen, Donoho and Saunders (1995) proposed to choose as the solution to this problem the expansion (11.85) that is defined by the coefficients that minimize the following functional

$$E(c) = \frac{1}{2} \|f(x) - \sum_{i=1}^n c_i \varphi_i(x)\|^2 + \gamma \sum_{i=1}^n |c_i| \quad (11.86)$$

where γ is some positive constant.

[†] As in the conditioning probability case to construct approximation (11.84) from the same data that was used for estimating the regression function, one can apply the leave-one-out technique.

In 1997 F. Girosi noted that in the case where there is no noise in the description of the function $f(x)$ the solution of a modified version of functional (11.86) is equivalent to the SV solution.

To describe this modification we have to remind the reader of some facts from the Theory of Reproducing Kernels Hilbert Spaces (RKHS).

11.12.1 Reproducing Kernels Hilbert Spaces

According to definition, a Hilbert space \mathcal{H} is a linear space where for any two elements $f_1(x)$ and $f_2(x)$ the value of the inner product $(f_1(x) * f_2(x))_{\mathcal{H}}$ is defined.

A reproducing kernels Hilbert space \mathcal{H} is a set of functions defined by the inner product and the kernel function $K(y, x)$ such that the following reproducing property

$$f(x) = (f(y) * K(y, x))_{\mathcal{H}} \quad \forall f(y) \in \mathcal{H},$$

holds true. According to the Mercer's theorem any positive definite function $K(y, x)$ defines the inner product in some Hilbert space and therefore, as we will see, defines some RKHS.

Indeed let

$$\phi_1(x), \dots, \phi_n(x), \dots$$

be the sequence of eigenfunctions for the kernel function $K(y, x)$ and

$$\lambda_1, \dots, \lambda_n, \dots$$

be the corresponding positive (since the kernel satisfies the Mercer condition) eigenvalues

$$\int K(y, x) \phi_k(y) dy = \lambda_k \phi_k(x).$$

The kernel $K(y, x)$ has the expansion

$$K(y, x) = \sum_{k=1}^{\infty} \lambda_k \phi_k(y) \phi_k(x). \quad (11.87)$$

It is easy to see that for Hilbert space \mathcal{H}

$$f(x) = \sum_{k=1}^{\infty} c_k \phi_k(x)$$

with the inner product

$$(f^*(x) * f^{**}(x))_{\mathcal{H}^*} = \sum_{k=1}^{\infty} \frac{c_k^* c_k^{**}}{\lambda_k} \quad (11.88)$$

the kernel (11.87) defines *RKHS*.

Indeed the following equalities are true

$$\begin{aligned} (f(x) * K(y, x)) &= \left(\sum_{k=1}^{\infty} c_k \phi_k(x) * \sum_{k=1}^{\infty} \lambda_k \phi_k(y) \phi_k(x) \right) \\ &= \sum_{k=1}^{\infty} \frac{\lambda_k c_k \phi_k(y)}{\lambda_k} = \sum_{k=1}^{\infty} c_k \phi_k(y) = f(y). \end{aligned}$$

11.12.2 Modified Sparse Approximation and its Relation to SV Machines

Let $K(x, y)$ be a reproducing kernel of a reproducing kernel Hilbert space \mathcal{H} and let x_1, \dots, x_ℓ be a set of points at which we know the values $y_i = f(x_i)$ of the target function $f(x) \in \mathcal{H}$. We make the following choice for the basis functions

$$\varphi_i(x) = K(x_i, x).$$

Our approximation function, therefore, is

$$f(x, c) = \sum_{i=1}^{\ell} c_i K(x_i, x).$$

Instead of minimizing functional (11.86), Chen et al. (1995) minimized the functional

$$E(c) = \frac{1}{2\ell} \sum_{j=1}^{\ell} (f(x_j) - \sum_{i=1}^n c_i \varphi_i(x_j))^2 + \gamma \sum_{i=1}^n |c_i|. \quad (11.89)$$

Girosi (1998) proposed to minimize the functional

$$G(c) = \frac{1}{2} \|f(x) - \sum_{i=1}^{\ell} c_i K(x_i, x)\|_{\mathcal{H}}^2 + \varepsilon \sum_{i=1}^{\ell} |c_i|. \quad (11.90)$$

We can expand this functional as follows:

$$\begin{aligned} G(c) &= \frac{1}{2} \|f\|_{\mathcal{H}}^2 - \sum_{i=1}^{\ell} c_i (f(x) * K(x_i, x))_{\mathcal{H}} \\ &\quad + \frac{1}{2} \sum_{i=1}^{\ell} c_i c_i (K(x_i, x) * K(x_i, x))_{\mathcal{H}} + \varepsilon \sum_{i=1}^{\ell} |c_i|. \end{aligned}$$

Using reproducing properties of kernel $K(x_i, x)$ we have

$$(f(x) * K(x_i, x))_{\mathcal{H}} = f(x_i) = y_i$$

$$(K(x_i, x) * K(x_j, x))_{\mathcal{H}} = K(x_i, x_j)$$

and therefore

$$G(c) = \frac{1}{2} \|f\|_{\mathcal{H}}^2 - \sum_{i=1}^{\ell} c_i y_i + \frac{1}{2} \sum_{i,j=1}^{\ell} c_i c_j K(x_j, x_j) + \varepsilon \sum_{i=1}^{\ell} |c_i|.$$

Introducing variables

$$c_i = \alpha_i^* - \alpha_i, \quad \alpha_i^*, \alpha_i \geq 0, \quad \alpha_i^* \alpha_i = 0, \quad i = 1, \dots, \ell$$

and disregarding the constant term $\frac{1}{2} \|f\|_{\mathcal{H}}^2$ we obtain the following method of solving of the sparse approximation problem: maximize the functional

$$W(\alpha, \alpha^*) = -\varepsilon \sum_{i=1}^{\ell} (\alpha_i^* + \alpha_i) + \sum_{i=1}^{\ell} (\alpha_i^* - \alpha_i) y_i - \frac{1}{2} \sum_{i,j=1}^{\ell} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) K(x_j, x_j)$$

subject to constraints

$$\alpha_i, \alpha_i^* \geq 0. \quad (11.91)$$

This solution of the problem of sparse function approximation coincides with the support vector solution if:

1. The function $f(x)$ is sufficiently smooth and there is no noise in its measurements. In this case the value C in the **SV** method can be chosen sufficiently large and constraints for **SV** method

$$0 \leq \alpha_i^*, \alpha_i \leq C$$

coincide with constraints (11.91).

2. One of the basis functions $\phi_0(x)$ is constant. In this case one does not need the additional constraint

$$\sum_{i=1}^{\ell} \alpha_i^* = \sum_{i=1}^{\ell} \alpha_i$$

that the **SV** method uses for choosing constant b .

Thus the **SV** method for function approximation that uses the linear ε -insensitive loss function provides sparse approximation of functions.

12

SV MACHINES FOR PATTERN RECOGNITION

This chapter considers the problem of digit recognition as an example of solving real-life pattern recognition problem using the SV machines. We show how to use SV machines to achieve high performance and discuss some ideas that can lead to performance increase.

12.1 THE QUADRATIC OPTIMIZATION PROBLEM

All experiments described in this chapter were conducted using SV machines constructed on the basis of quadratic optimization techniques for the soft margin objective function. The main element of the corresponding algorithms is constructing the optimal separating hyperplane. To construct the optimal hyperplane for the pattern recognition problem, we maximize the quadratic form

$$W(\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j K(x_i, x_j) y_i y_j \quad (12.1)$$

subject to constraints

$$\begin{aligned} \sum_{i=1}^{\ell} y_i \alpha_i &= 0, \\ 0 \leq \alpha_i &\leq C, \quad i = 1, 2, \dots, \ell. \end{aligned} \quad (12.2)$$

To estimate a functional dependency in the sets of real-valued functions,

we maximize the slightly different functional

$$W(\alpha) = -\sum_{i=1}^{\ell} \varepsilon_i (\alpha_i^* + \alpha_i) + \sum_{i=1}^{\ell} (\alpha_i^* - \alpha_i) y_i - \frac{1}{2} \sum_{i,j=1}^{\ell} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) K(x_i, x_j) \quad (12.3)$$

subject to constraint

$$\begin{aligned} \sum_{i=1}^{\ell} \alpha_i &= \sum_{i=1}^{\ell} \alpha_i^*, \\ 0 \leq \alpha_i &\leq C, \quad i = 1, 2, \dots, \ell, \\ 0 \leq \alpha_i^* &\leq C^*, \quad i = 1, 2, \dots, \ell, \end{aligned} \quad (12.4)$$

where the kernel $K(x_i, x_j)$ satisfies Mercer's condition (see Chapter 10, Section 10).

The methods for solving these two optimization problems are identical. Therefore we consider only methods for the pattern recognition problem.

12.1.1 Iterative Procedure for Specifying Support Vectors

The goal is to construct optimization algorithms that are capable of using hundreds of thousands of observations and construct decision rules based on tens of thousands of support vectors.

To find the maximum of the functional $W(\alpha)$ in a such high-dimensional space, one has to take into account that the solution α^* to our optimization problem is an ℓ -dimensional vector where only a small number of coordinates (the ones that correspond to support vectors) are not equal to zero.

Therefore one iteratively maximizes the objective function in the different subspaces where coordinates are nonzero using the following algorithm:

1. At the first iteration (arbitrarily) assign most of the coordinates to zero (let us call them the nonactive variables) and look for conditional maximum with respect to the remaining coordinates (active variables). Therefore one considers a reduced optimization problem where the functional (12.1) has a reasonably small number of active variables (say, several hundreds).
2. Let vector $\alpha(1)$ be the solution to this reduced optimization problem and let $W(\alpha(1))$ be a corresponding value of the functional. Check whether the vector $\alpha(1)$ defines the solution to the desired optimization problem. To be a solution to the desired optimization problem the ℓ -dimensional vector $\alpha^*(1) = (\alpha_1^*(1), \dots, \alpha_\ell^*(1))$ (most coordinates of which are equal to zero) must satisfy the conditions

$$\begin{aligned}
 y_k(\sum_{i=1}^{\ell} \alpha_i^0 y_i K(x_i, x_k) + b) &= 1 && \text{if } 0 < \alpha_i^0(1) < C, \\
 y_k(\sum_{i=1}^{\ell} \alpha_i^0 y_i K(x_i, x_k) + b) &\geq 1 && \text{if } \alpha_i^0(1) = 0, \\
 y_k(\sum_{i=1}^{\ell} \alpha_i^0 y_i K(x_i, x_k) + b) &\leq 1 && \text{if } \alpha_i^0(1) = C, \quad i = 1, \dots, \ell.
 \end{aligned} \tag{12.5}$$

If conditions (12.5) are satisfied, then one has constructed the desired approximation.

3. Suppose that for some i the conditions (12.5) fail. Then construct the next approximation $\alpha(2)$. For this purpose make nonactive (by assigning to zero) those variables for which $\alpha_i(1) = 0$ and make active some number of variables α_i corresponding to x_i for which the inequality constraints (12.5) do not hold.

In this new space maximize the reduced quadratic form. Start maximization with the initial conditions

$$\alpha_i^{in}(2) = \begin{cases} \alpha_i(1) & \text{if } \alpha_i(1) \neq 0, \\ 0 & \text{otherwise.} \end{cases} \tag{12.6}$$

Since

$$W(\alpha^{in}(2)) = W(\alpha(1)),$$

the maximum for the optimal solution in the second iteration exceeds the maximum of the optimal solution for the first iteration.

4. Continue these iterations until the maximum is approached (satisfy the condition (12.5)).

The described procedure works well until the number of support vectors is less than several thousand.

To obtain the solution with a large number of support vectors (up to hundreds of thousands), Osuna, Freund, and Girosi (1997a,b) suggested the following procedure:

1. Arbitrarily choose $|B|$ points from the data set.
2. Solve the optimization problem defined by variables in B .
3. While there exist some points in the training set for which the conditions (12.5) are not valid, replace any points and corresponding weights a from the set B with these points and corresponding weights a and solve the new optimization problem with respect to variables in a new set B , keeping fixed coefficients a corresponding to points that do not belong to set B .

Since the algorithm strictly improves the objective function at any iteration, it will not cycle. Since the objective function is bounded ($W(\alpha)$ is convex quadratic and the feasible region is bounded), the algorithm must converge to global optimal solution in a finite number of iterations. Platt (1998) and

Joachims (1988) suggested modifications of this procedure that speed up the learning process for large databases.

12.1.2 Methods for Solving the Reduced Optimization Problem

There are a number of methods for solving quadratic optimization problems. However, we need to solve a special (simple) quadratic optimization problem that is described by one constraint of equality type and coordinate constraints (12.2) (box constraints). For this specific constraint, one can construct special optimization methods that are more efficient than general quadratic optimization methods. For example, one can construct methods based on the conjugate gradient procedure, the interior point method, and the projection procedure. There exist standard packages implementing these methods. Any of these can be used for constructing an SV machine. Below we describe experiments with SV machines that were conducted using MINOS 5.4, LOGO, and IQP.

12.2 DIGIT RECOGNITION PROBLEM. THE U.S. POSTAL SERVICE DATABASE

Since the first experiments of Rosenblatt, the interest in the problem of learning to recognize handwritten digits has remained strong. In the following we describe the results of experiments on learning the recognition of handwritten digits using different SV machines. We also compare the SV machine results to results obtained by other classifiers. The experiments were conducted using two different databases: the US Postal Service database and the National Institute of Standard and Technology (NIST) database.

In this section we describe experiments with the U.S. Postal Service database, and in the next section we describe experiments with the NIST database.

12.2.1 Performance for the U.S. Postal Service Database

The U.S. Postal Service database contains 7291 training patterns and 2007 test patterns collected from real-life zip codes. The resolution of the database is 16 x 16 pixel, and therefore the dimensionality of the input space is 256. Figure 12.1 gives examples from this database.

Table 12.1 describes the performance of various classifiers, solving this problem.[†]

For constructing the decision rules, three types of SV machines were used[‡]:

[†]The results of human performance were reported by J. Bromley and E. Sackinger; the results of C4.5 were obtained by C. Cortes; the results for the two layer neural net were obtained by B. Scholkopf; the results for the special-purpose neural network architecture with five layers (LeNet I) were obtained by Y. LeCun et al.

[‡]The results were obtained by C. Burges, C. Cortes, and B. Scholkopf.

65510189101037559991103255
 0345805401054530544605853
 0540205091373973731137816
 60403113837962240000=7873
 720-722047014890025080875
 4597038973029411194107691
 6068009680096812711130571
 0557102997095870173159740
 0296712967139673496816664
 41735767230871048212919+1
 7334964814148080484334876
 8482074700648240058542501
 4865514227442544720348134
 6419920874443221440774558
 4447114423044258455014209
 5440153538544095340-03054
 1410304441967203306833502
 3333914130551825036700055
 0340641338411264141661026
 4111662156604367039841218

8388 8381 8382 8383 8384 8385 8386 8387 8388 8389 8310 8311 8312 8313 8314 8315 8316 8317 8318 8319 8320 8321 8322 8323 8324
 8325 8326 8327 8328 8329 8330 8331 8332 8333 8334 8335 8336 8337 8338 8339 8340 8341 8342 8343 8344 8345 8346 8347 8348 8349
 8350 8351 8352 8353 8354 8355 8356 8357 8358 8359 8360 8361 8362 8363 8364 8365 8366 8367 8368 8369 8370 8371 8372 8373 8374
 8375 8376 8377 8378 8379 8380 8381 8382 8383 8384 8385 8386 8387 8388 8389 8390 8391 8392 8393 8394 8395 8396 8397 8398 8399
 8400 8401 8402 8403 8404 8405 8406 8407 8408 8409 8410 8411 8412 8413 8414 8415 8416 8417 8418 8419 8420 8421 8422 8423 8424
 8425 8426 8427 8428 8429 8430 8431 8432 8433 8434 8435 8436 8437 8438 8439 8440 8441 8442 8443 8444 8445 8446 8447 8448 8449
 8450 8451 8452 8453 8454 8455 8456 8457 8458 8459 8460 8461 8462 8463 8464 8465 8466 8467 8468 8469 8470 8471 8472 8473 8474
 8475 8476 8477 8478 8479 8480 8481 8482 8483 8484 8485 8486 8487 8488 8489 8490 8491 8492 8493 8494 8495 8496 8497 8498 8499
 8500 8501 8502 8503 8504 8505 8506 8507 8508 8509 8510 8511 8512 8513 8514 8515 8516 8517 8518 8519 8520 8521 8522 8523 8524
 8525 8526 8527 8528 8529 8530 8531 8532 8533 8534 8535 8536 8537 8538 8539 8540 8541 8542 8543 8544 8545 8546 8547 8548 8549
 8550 8551 8552 8553 8554 8555 8556 8557 8558 8559 8560 8561 8562 8563 8564 8565 8566 8567 8568 8569 8570 8571 8572 8573 8574
 8575 8576 8577 8578 8579 8580 8581 8582 8583 8584 8585 8586 8587 8588 8589 8590 8591 8592 8593 8594 8595 8596 8597 8598 8599
 8600 8601 8602 8603 8604 8605 8606 8607 8608 8609 8610 8611 8612 8613 8614 8615 8616 8617 8618 8619 8620 8621 8622 8623 8624
 8625 8626 8627 8628 8629 8630 8631 8632 8633 8634 8635 8636 8637 8638 8639 8640 8641 8642 8643 8644 8645 8646 8647 8648 8649
 8650 8651 8652 8653 8654 8655 8656 8657 8658 8659 8660 8661 8662 8663 8664 8665 8666 8667 8668 8669 8670 8671 8672 8673 8674
 8675 8676 8677 8678 8679 8680 8681 8682 8683 8684 8685 8686 8687 8688 8689 8690 8691 8692 8693 8694 8695 8696 8697 8698 8699
 8700 8701 8702 8703 8704 8705 8706 8707 8708 8709 8710 8711 8712 8713 8714 8715 8716 8717 8718 8719 8720 8721 8722 8723 8724
 8725 8726 8727 8728 8729 8730 8731 8732 8733 8734 8735 8736 8737 8738 8739 8740 8741 8742 8743 8744 8745 8746 8747 8748 8749
 8750 8751 8752 8753 8754 8755 8756 8757 8758 8759 8760 8761 8762 8763 8764 8765 8766 8767 8768 8769 8770 8771 8772 8773 8774
 8775 8776 8777 8778 8779 8780 8781 8782 8783 8784 8785 8786 8787 8788 8789 8790 8791 8792 8793 8794 8795 8796 8797 8798 8799

FIGURE 12.1. Examples of patterns (with labels) from the U.S. Postal Service database.

Table 12.1. Human performance and performance of the various learning machines, solving the problem of digit recognition on U.S. Postal Service data

Classifier	Raw error%
Human performance	2.5
Decision tree, C4.5	16.2
Best two-layer neural network	5.9
Five-layer network (LeNet 1)	5.1

1. A polynomial machine with kernel function:

$$K(x, x_i) = \left(\frac{(x * x_i)}{256} \right)^d, \quad d = 1, \dots, 7.$$

2. A radial basis function machine with kernel function:

$$K(x, x_i) = \exp \left\{ - \frac{|x - x_i|^2}{256\sigma^2} \right\}$$

3. A two-layer neural network machine with kernel function:

$$K(x, x_i) = \tanh \left(\frac{b(x * x_i)}{256} - c \right),$$

where

$$\tanh(u) = \frac{e^u - e^{-u}}{e^u + e^{-u}}$$

All machines constructed 10 classifiers, each one separating one class from the rest. The 10-class classification was done by choosing the class with the largest output value.

The results of these experiments are given in Tables 12.2, 12.3, and 12.4. For different types of **SV** machines, the tables show the parameters for the machines, the corresponding performance, and the average (over one classifier) number of support vectors.

Note that for this problem, all types of **SV** machines demonstrate approximately the same performance. This performance is better than the perfor-

Table 12.2. Results of digit recognition experiments with polynomial SV machines (with the inner products $((x * y)/256)^{\text{degree}}$)

Degree	1	2	3	4	5	6
Raw error:	8.9	4.7	4.0	4.2	4.5	4.5
Average number of SV:	282	237	274	321	374	422

Table 12.3. Results of digit recognition experiments with RBF SV machines (with inner products $\exp\{-||x - y||^2/256\sigma^2\}$)

σ :	4.0	1.5	0.30	0.25	0.2	0.1
Raw error:	5.3	4.9	4.2	4.3	4.5	4.6
Average number of SV:	266	237	274	321	374	422

Table 12.4. Results of digit recognition experiments with NN SV machines (with inner products $1.04 \tanh\{2(x \cdot y)/256 - \theta\}$)

θ :	0.8	0.9	1.0	1.2	1.3	1.4
Raw error:	6.3	4.9	4.2	4.3	4.5	4.6
Average number of SV:	206	237	274	321	374	422

Table 12.5. The total number (in 10 classifiers) of support vectors for various SV machines and percentage of common support vectors

	Poly	RBF	NN	Common
Total number of support vectors:	1677	1727	1611	1377
Percentage of common support vectors:	82	80	85	100

mance of any other type of learning machine solving the digit recognition problem by constructing the decision rules on the basis of the entire U.S. Postal Service database.[†]

In these experiments, one important feature was observed: Different types of SV machine use approximately the same set of support vectors. The percentage of common support vectors for three different classifiers exceeded 80%.

Table 12.5 describes the total number of different support vectors for 10 classifiers of different machines: polynomial machine (Poly), radial basis function machine (RBF), and neural network machine (NN). It shows also the number of common support vectors for all machines.

Table 12.6 describes the percentage of support vectors of the classifier given in the columns contained in the support vectors of the classifier given in the rows.

[†]Note that by using a local approximation approach described in Section 5.7 (that does not construct entire decision rule but approximates the decision rule at any point of interest), one can obtain a better result: 3.3% error rate (Bottou and Vapnik, 1992). The best result for this database, 2.7%, was obtained by Simard, LeCun, and Denker (1993) without using any learning methods. They suggested a special method of elastic matching with 7200 templates using a smart concept of distance (so-called tangent distance) that takes into account invariance with respect to small translations, rotations, distortions, and so on (Simard, LeCun, and Denker, 1993). We will discuss this method in Section 12.4.

Table 12.6. Percentage of common (total) support vectors for two SV machines

	Poly	RBF	NN
Poly	100	84	94
RBF	87	100	88
NN	91	82	100

12.2.2 Some Important Details

In this subsection we give some important details on solving the digit recognition problem using a polynomial SV machine.

The training data are not linearly separable. The total number of misclassifications on the training set for linear rules is equal to 340 ($\approx 5\%$ errors). For second-degree polynomial classifiers the total number of mis-classifications on the training set is down to four. These four misclassified examples (with desired labels) are shown in Fig. 12.2. Starting with polynomials of degree three, the training data are separable.

Table 12.7 describes the results of experiments using decision polynomials (10 polynomials, one per classifier in one experiment) of various degrees. The number of support vectors shown in the table is a mean value per classifier.

Note that the number of support vectors increases slowly with the degree of the polynomial. The seventh-degree polynomial has only 50% more support vectors than the third-degree polynomial.[†]

The dimensionality of the feature space for a seventh-degree polynomial is, however, 10^{10} times larger than the dimensionality of the feature space for a third-degree polynomial classifier. Note that the performance does not change significantly with increasing dimensionality of the space—indicating no overfitting problems.

To choose the degree of the best polynomial for one specific classifier we estimate the VC dimension (using the estimate $D_\ell^2 |w_\ell|^2$, see Chapter 10, Section 10.7) for all constructed polynomials (from degree two up to degree seven) and choose the one with the smallest estimate of the VC dimension. In this way we found the 10 best classifiers (with different degrees of polynomials) for the 10 two-class problems. These estimates are shown

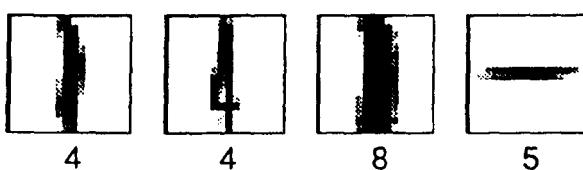


FIGURE 12.2. Labeled examples of training errors for the second-degree polynomials.

[†]The relatively high number of support vectors for the linear function is due to nonseparability: The number 282 includes both support vectors and misclassified data.

Table 12.7. Results of experiments with polynomials of the different degrees

Degree of Polynomial	Dimensionality of Feature space	Support Vectors	Raw Error
1	256	282	8.9
2	$\sim 33,000$	227	4.7
3	$\sim 1 \times 10^6$	274	4.0
4	$\sim 1 \times 10^9$	321	4.2
5	$\sim 1 \times 10^{12}$	374	4.3
6	$\sim 1 \times 10^{14}$	377	4.5
7	$\sim 1 \times 10^{16}$	422	4.5

on Fig. 12.3, where for all 10 two-class decision rules, the estimated VC dimension is plotted versus the degree of the polynomials.

The question is, *Do the polynomials with the smallest estimate of the VC dimension provide the best classifier?* To answer this question we constructed Table 12.8, which describes the performance of the classifiers for each degree of polynomial.

Each row describes one two-class classifier separating one *digit* (stated in the first column) from the all other digits.

The remaining columns contain:

deg.: the degree of the polynomial as chosen (from two up to seven) by the described procedure,

dim.: the dimensionality of the corresponding feature space, which is also the maximum possible VC dimension for linear classifiers in that space,

h_{est.}: the VC dimension estimate for the chosen polynomial (which is much smaller than the number of free parameters),

Number of test errors: the number of test errors, using the constructed polynomial of corresponding degree; the boxes show the number of errors for the chosen polynomial.

Thus, Table 12.7 shows that for the SV polynomial machine there are no overfitting problems with increasing degree of polynomials, while Table 12.8 shows that even in situations where the difference between the best and the worst solutions is small (for polynomials starting from degree two up to degree seven), the theory gives a method for approximating the best solutions (finding the best degree of the polynomial).

Note also that Table 12.8 demonstrates that the problem is essentially non-linear. The difference in the number of errors between the best polynomial classifier and the linear classifier can be as much as a factor of four (for digit 9).

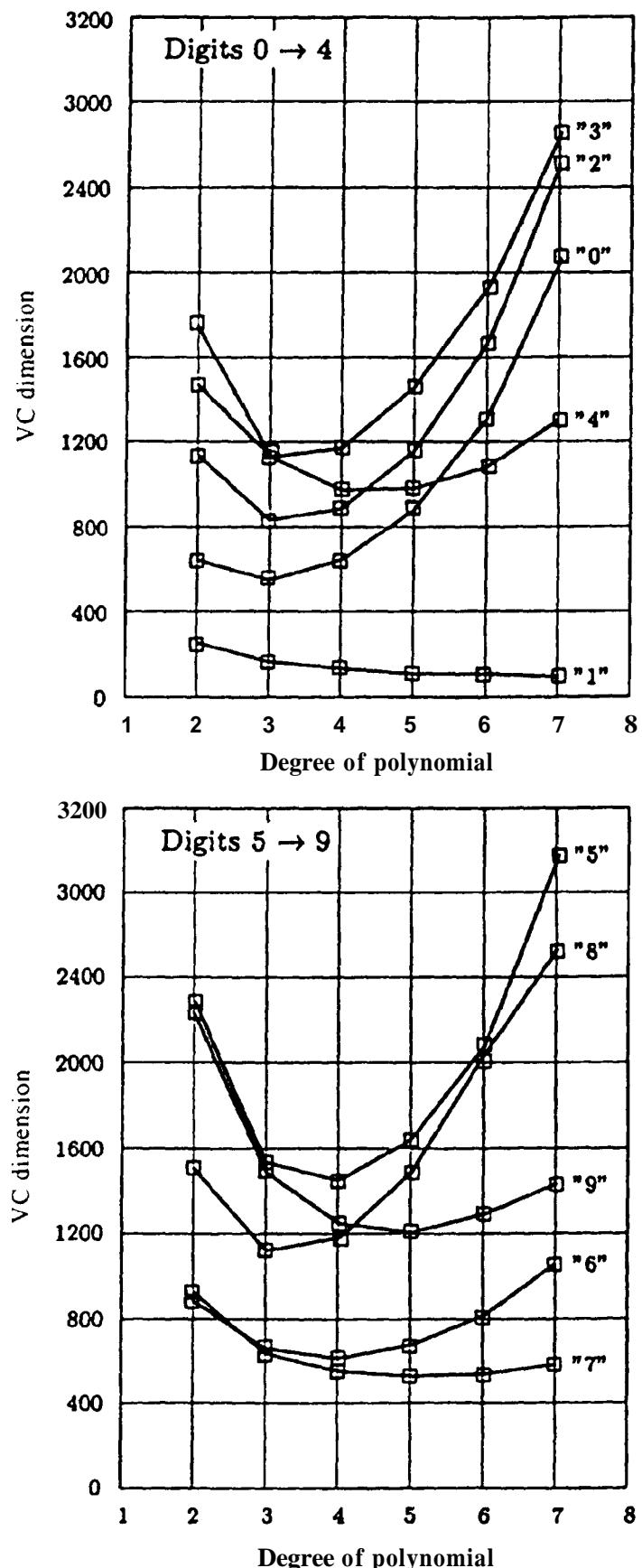


FIGURE 12.3. The estimate of the VC dimension of the best element of the structure defined by the value $D_i^2 |w_i|^2$ versus the degree of polynomial for various two-class digit recognition problems.

Table 12.8. Experiments on choosing the best degree of polynomial"

Digit	Chosen Classifier			Number of Test Errors						
	deg.	dim.	$h_{\text{est.}}$	1	2	3	4	5	6	7
0	3	$\sim 10^6$	530	36	14	11	11	12	17	
1	7	$\sim 10^{16}$	101	17	15	14	11	10	10	10
2	3	$\sim 10^6$	842	53	32	28	26	28	27	32
3	3	$\sim 10^6$	1157	57	25	22	22	22	22	23
4	4	$\sim 10^9$	962	50	32	32	30	30	29	33
5	3	$\sim 10^6$	1090	37	20	22	24	24	26	28
6	4	$\sim 10^9$	626	23	12	12	15	17	17	19
7	5	$\sim 10^{12}$	530	25	15	12	10	11	13	14
8	4	$\sim 10^9$	1445	71	33	28	24	28	32	34
9	5	$\sim 10^{12}$	1226	51	18	15	11	11	12	15

^aThe boxes indicate the chosen order of a polynomial.

12.2.3 Comparison of Performance of the SV Machine with Gaussian Kernel to the Gaussian RBF Network

Since the RBF network with a Gaussian kernel produces the same type of decision rules

$$f_{\text{RBF}}(x) = \text{sign} \left(\sum_{k=1}^N a_k \exp\{-||x - c_k||^2/\sigma^2\} + b \right)$$

that is produced by the SV machine

$$f_{\text{SV}}(x) = \text{sign} \left(\sum_{k=1}^N y_i \alpha_k \exp\{-||x - x_k||^2/\sigma^2\} + b \right)$$

but uses completely different ideas for choosing the parameters of decision rules (the centers c_k instead of support vectors x_k and coefficients of expansion a_k that minimize mean square deviation instead of coefficients α_k that make an optimal separating hyperplane), it is important to compare their performance.

The result of this comparison should answer the following questions:

Is it true that support vectors are the best choice for placing the centers?

Is it true that the SV estimate of expansion coefficients is the best?

To answer these questions the RBF networks were compared with SV machines on the problem of U.S. Postal Service digit recognition.[†]

The classical RBF method does not specify how to choose the number of centers and the parameter σ . Therefore in these experiments the same parameter σ was used. Also for RBF networks the number of centers was chosen equal to the number of support vectors that were used by the SV machine (the variation in the number of centers did not improve the performance of the RBF network).

In the first experiment a classical RBF network was used, which defined centers by k-means clustering and constructed weights by error back-propagation. The obtained the performance was a 6.7% error rate.

In the second experiment the support vectors were used as centers and weights were chosen by error back-propagation. In this experiment we obtained 4.9% error rate. The performance of the SV machine is a 4.2% error rate.

The result of this experiment are summarized in Table 12.9.

To understand the geometry of this experiment better, let us compare the RBF solution to the SV solution of the simple two-dimensional classification problem given in Fig. 12.4: Find a decision function separating balls from circles. Solving this problem the SV machine chooses five support vectors, two for balls and three for circles (they are indicated by extra circles). The five centers that were chosen in the RBF network using the k-means method

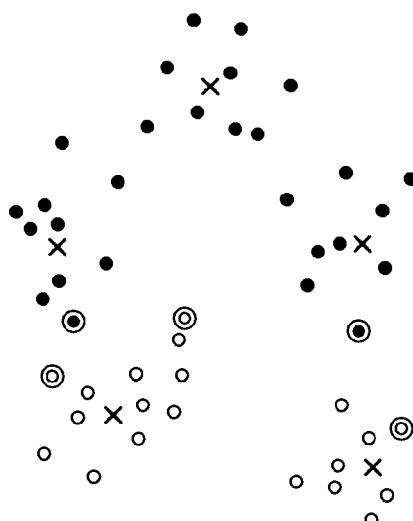


FIGURE 12.4. The support vectors (indicated by extra circles) and RBF centers (indicated by crosses) for simple classification problem.

[†]The experiments were conducted in the AI laboratory at MIT. See B. Scholkopf et al. (1997).

Table 12.9. Results of digit recognition experiments with three networks: 1) classical RBF networks, 2) hybrid networks with SV centers and the classical method for choosing weights, and 3) the SV machine

	RBF Networks	SV Centers	SV Machine
Training error	1.7%	0.0%	0.0%
Test error	6.7%	4.9%	4.2%

are indicated by crosses (three for balls and two for circles). Note that in contrast to the RBF centers the support vectors are chosen with respect to the classification task to be solved.

12.2.4 The Best Results for U.S. Postal Service Database

In the previous section, in describing the best results for solving the digit recognition problem using the U.S. Postal Service database by constructing an entire (not local) decision rule we gave two figures:

- 5.1% error rate for the neural network LeNet 1
- 4.0% error rate for a polynomial SV machine

However, the best results achieved for this database are:

- 3.3% error rate for the local learning approach, described in Chapter 6, Section 6.6
- 2.9% error rate for a sparse polynomial of degree 4 ($d_1 = 2$, $d_2 = 2$) SV machine (which will be described[†] in Section 12.5) and the record
- 2.7% error rate for tangent distance matching to templates given by the training set

Therefore the best results for the U.S. postal service database (2.7% of error rate) was achieved without any learning procedure, using one nearest-neighbor algorithm but using important a priori information about invariants of handwritten digits incorporated into special measure of distance between two vectors, the so-called tangent distance.

The main lesson that one has to learn from this fact is that when one has a relatively small amount of training examples, the effect of using a priori information can be even more significant than the effect of using a learning machine with a good generalization ability.

In the next section we present an example that shows that this is not true when the number of training data is large. However, in all cases to achieve the best performances, one must take into account the available a priori information.

[†]This result was obtained by B. Scholkopf.

12.3 TANGENT DISTANCE

In 1993 Simard et al. suggested that we use the following a priori information about handwritten digits[†]:

A reasonably small continuous transformation of a digit does not change its class.

This observation has the following mathematical expression. Consider the plane defined by the pair (t,s) . The following equation describes the general form of linear transformation of a point in the plane:

$$\begin{bmatrix} t^* \\ s^* \end{bmatrix} = \begin{vmatrix} 1+a & b \\ c & 1+d \end{vmatrix} \begin{bmatrix} t \\ s \end{bmatrix} + \begin{bmatrix} e \\ f \end{bmatrix}$$

This transformation is defined by six independent parameters. Consider the following expansion of this transformation into six basic transformations, each of which is defined by one parameter:

1. *Horizontal Translation.* The case where

$$a = b = c = d = f = 0.$$

Horizontal translation is described by equations

$$\begin{aligned} t^* &= t + e, \\ s^* &= s. \end{aligned} \tag{12.7}$$

2. *Vertical Translation.* The case where

$$a = b = c = d = e = 0.$$

Vertical translation is described by equations

$$\begin{aligned} t^* &= t, \\ s^* &= s + f. \end{aligned} \tag{12.8}$$

3. *Rotation.* The case where

$$a = d = e = f = 0, \quad b = -c.$$

Rotation is described by equations

$$\begin{aligned} t^* &= t + bs, \\ s^* &= s - bt. \end{aligned} \tag{12.9}$$

[†]This observation is correct for many different type of images, not only for digits.

4. Scaling. The case where

$$c = b = e = f = 0, \quad a = d$$

Scaling is described by equations

$$\begin{aligned} t^* &= t + at, \\ s^* &= s + as. \end{aligned} \tag{12.10}$$

5. Axis Deformation. The case where

$$a = d = e = f = 0, \quad b = c.$$

Axis deformation is described by equations

$$\begin{aligned} t^* &= t + cs, \\ s^* &= ct + s. \end{aligned} \tag{12.11}$$

6. Diagonal Deformation. The case where

$$b = c = e = f = 0, \quad a = -d.$$

Diagonal deformation is described by equations

$$\begin{aligned} t^* &= t + dt, \\ s^* &= s - ds. \end{aligned} \tag{12.12}$$

It is easy to check that any linear transformation can be obtained combining these six basic transformations.

Now let a continuous function $x(t, s)$ be defined on the plane (t, s) . Using basic transformations we now define the following six functions in the plane, which are called Lie derivatives of the function $x(t, s)$:

1. Function $x^{(1)}(t, s)$, which for any point of the plane defines the rate of change $x(t, s)$ in the horizontal translation direction:

$$x^{(1)}(t, s) = \lim_{e \rightarrow 0} \frac{x(t + e, s) - x(t, s)}{e} \tag{12.13}$$

2. Function $x^{(2)}(t, s)$, which for any point of the plane defines the rate of change $x(t, s)$ in the vertical translation direction:

$$x^{(2)}(t, s) = \lim_{f \rightarrow 0} \frac{x(t, s + f) - x(t, s)}{f} \tag{12.14}$$

3. Function $x^{(3)}(t, s)$, which for any point of the plane defines the rate of change $x(t, s)$ in the rotation direction:

$$\begin{aligned} x^{(3)}(t, s) &= \lim_{b \rightarrow 0} \frac{x(t + bs, s - bt) - x(t, s)}{b} \\ &= \lim_{b \rightarrow 0} \frac{x(t + bs, s - bt) - x(t, s - bt)}{b} + \lim_{b \rightarrow 0} \frac{x(t, s - bt) - x(t, s)}{b} \\ &= sx^{(1)}(t, s) - tx^{(2)}(t, s). \end{aligned} \quad (12.15)$$

4. Function $x^{(4)}(t, s)$, which for any point of the plane defines the rate of change $x(t, s)$ in the scaling direction:

$$x^{(4)}(t, s) = \lim_{a \rightarrow 0} \frac{x(t + at, s + as) - x(t, s)}{a} = tx^{(1)}(t, s) + sx^{(2)}(t, s). \quad (12.16)$$

5. Function $x^{(5)}(t, s)$, which for any point of the plane defines the rate of change $x(t, s)$ in the axis deformation direction:

$$x^{(5)}(t, s) = \lim_{c \rightarrow 0} \frac{x(t + cs, s + ct) - x(t, s)}{c} = sx^{(1)}(t, s) + tx^{(2)}(t, s). \quad (12.17)$$

6. Function $x^{(6)}(t, s)$, which for any point of the plane defines the rate of change $x(t, s)$ in diagonal deformation direction:

$$x^{(6)}(t, s) = \lim_{d \rightarrow 0} \frac{x(t + dt, s - ds) - x(t, s)}{d} = tx^{(1)}(t, s) - sx^{(2)}(t, s) \quad (12.18)$$

Along with six classical functions that define small linear transformation of the function, I? Simard suggested that we use the following function, which is responsible for *thickness deformation*:

7. This function is defined in any point of the plane as follows:

$$x^{(7)}(t, s) = \left(x^{(1)}(t, s) \right)^2 + \left(x^{(2)}(t, s) \right)^2. \quad (12.19)$$

All seven functions can be easily calculated for any smooth continuous function $x(t, s)$. To define them it is sufficient to define the first two functions $x^{(1)}(t, s)$ and $x^{(2)}(s, t)$; the rest of the five functions can be calculated using these two.

These functions are used to describe small transformation of the function $x(t, s)$:

$$x^*(t, s) = x(t, s) + \sum_{i=1}^7 a_i x^{(i)}(t, s), \quad (12.20)$$

where parameters that specify the transformation are small $|a| \leq C$.

Note that Lie derivatives are defined for continuous functions $x(t, s)$. In the problem of digit recognition, however, functions are described by their values in the pixels $\hat{x}(t', s')$, $t', s' = 1, 2, \dots, 2^k$. They are discrete. To be able to use methods based on the theory of small transformations for smooth functions, one has to first approximate the discrete functions by smooth functions. This can be done in many ways, for example, by convolving a discontinuous function with a Gaussian—in other words, by smoothing discontinuous functions as follows:

$$x(s, t) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \hat{x}(t', s') \exp \left\{ -\frac{(s - s')^2 + (t - t')^2}{2\sigma^2} \right\} dt' ds'.$$

Examples of the smoothed function $x(t, s)$, its six Lie derivatives, and Simard's thickness deformation function are shown in Fig. 12.5.

Figure 12.6 shows the original image and the new images obtained using linear transformation (12.20) with various coefficients a_i , $i = 1, \dots, 7$.

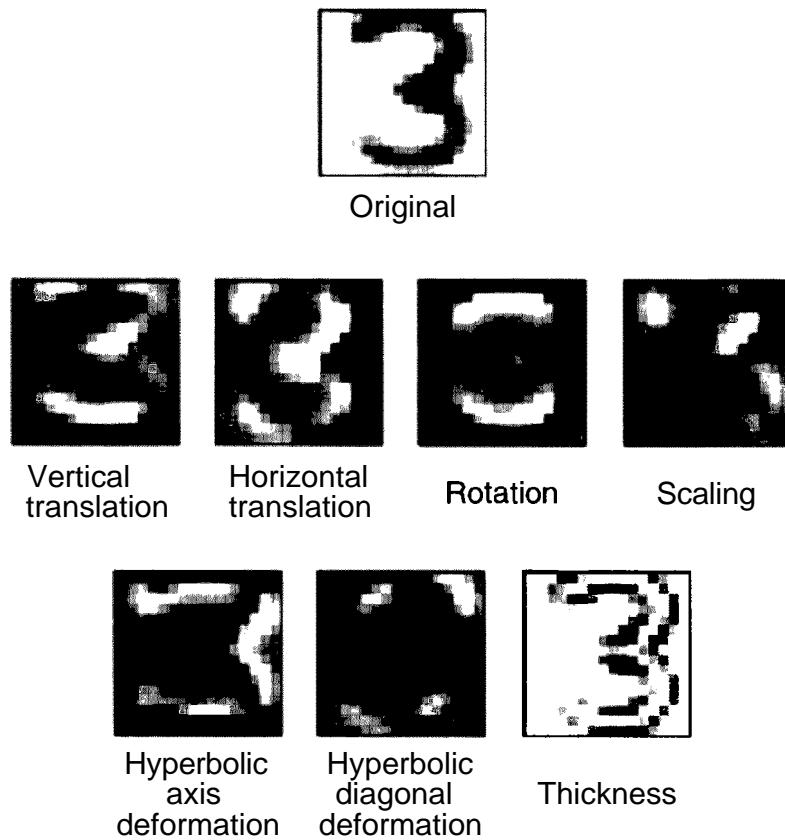


FIGURE 12.5. Smoothed image and calculated functions $x^{(i)}(t, s)$, $i = 1, \dots, 7$.

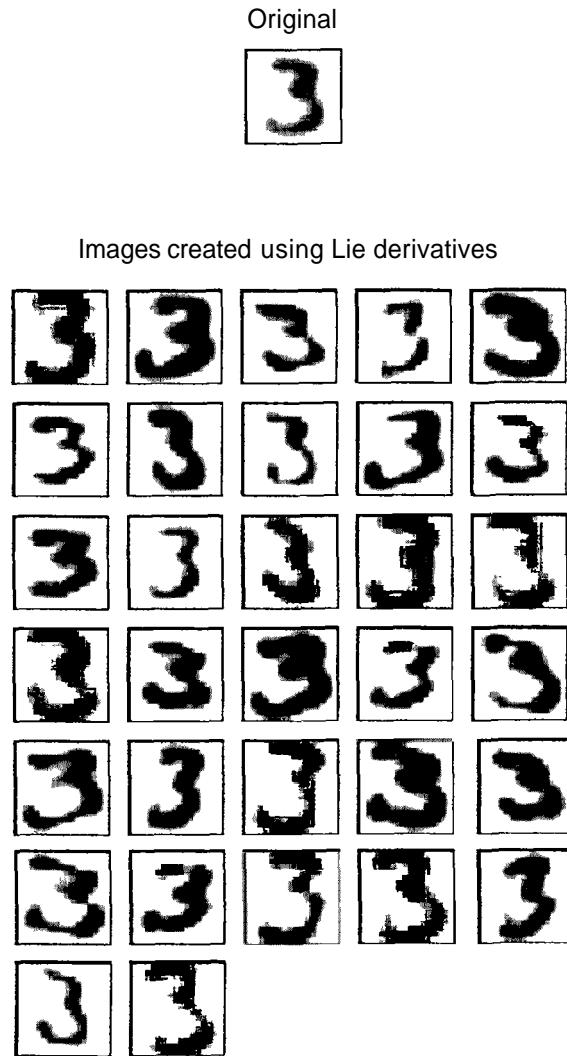


FIGURE 12.6. Digits obtained from one given example using linear transformation (12.20) with various coefficients.

Now we are ready to define the following measure of difference between two images $x_1(t, s)$ and $x_2(t, s)$:

$$\begin{aligned} & \rho^2(x_1(t, s), x_2(t, s)) \\ &= \min_{a, b} \left(x_1(t, s) + \sum_{i=1}^7 a_i x_1^{(i)}(t, s) - x_2(t, s) - \sum_{i=1}^7 b_i x_2^{(i)}(t, s) \right)^2. \end{aligned} \quad (12.21)$$

This measure defines distortion after invariant matching of two images. It was called the *tangent distance*.[†]

Using tangent distance in the one nearest-neighbor algorithm and 7300 training examples as templates, the record for the U.S. Postal Service database was achieved.

[†]From a formal point of view, expression (12.21) does not define distance, since it does not satisfy the triangle inequality.

12.4 DIGIT RECOGNITION PROBLEM: THE NIST DATABASE

12.4.1 Performance for NIST Database

In 1993, responding to the community's need for benchmarking, the U.S. National Institute of Standard and Technology (NIST) provided a database of handwritten characters containing 60,000 training images and 10,000 test data, in which characters are described as vectors in $20 \times 20 = 400$ pixel space.

For this database a special neural network (LeNet 4) was designed. The following is how the article reporting the benchmark studies (Bottou et al., 1994) describes the construction of LeNet 4:

For quite a long time, LeNet 1 was considered state of the art. The local learning classifier, the SV classifier, and tangent distance classifier were developed to improve upon LeNet 1—and they succeeded in that. However, they in turn motivated a search for an improved neural network architecture. This search was guided in part by estimates of the capacity of various learning machines, derived from measurements of the training and test error (on the large NIST database) as a function of the number of training examples. We discovered that more capacity was needed. Through a series of experiments in architecture, combined with an analysis of the characteristics of recognition errors, the five-layer network LeNet 4 was crafted.

In these benchmarks, two learning machines that construct entire decision rules—(1) LeNet 4 and (2) Polynomial SV machine (polynomial of degree four)—provided the same performance: 1.1% test error.[†]

The local learning approach and tangent distance matching to 60,000 templates also yield the same performance: 1.1% test error.

Recall that for a small (U.S. Postal Service) database the best result (by far) was obtained by the tangent distance matching method that uses a priori information about the problem (incorporated in the concept of tangent distance). As the number of examples increases to 60,000, the advantage of a priori knowldgc dcrcascd. The advantage of the local learning approach also decreased with the increasing number of observations.

LeNet 4, crafted for the NIST database, demonstrated remarkable improvement in performance when compared to LeNet 1 (which has 1.7% test errors for the NIST database[‡]).

The standard polynomial SV machine also performed well. We continue the quotation (Bottou et al., 1994):

[†] Unfortunately, one cannot compare these results to the results described in Section 12.2. The digits from the NIST database are different from the U.S. Postal Service database.

[‡] Note that LeNet 4 has an advantage for the large (60,000 training examples) NIST database. For a small (U.S. Postal Service) database containing 7000 training examples, the network with smaller capacity, LeNet 1, is better.

The SV machine has excellent accuracy, which is most remarkable, because unlike the other high-performance classifiers it *does not include knowledge about the geometry of the problem*. In fact this classifier would do just as well if the image pixel were encrypted, for example, by a fixed random permutation.

12.4.2 Further Improvement

Further improvement of results for the NIST database was obtained both for neural networks and for SV machines.

A new neural network was created—the so-called LeNet 5—that has a five-layer architecture similar to LeNet 4, but more feature maps, and a larger fully connected layer. LeNet 5 has 60,000 free parameters (LeNet 4 has 17,000), most of them in the last two layers.

It is important to notice that LeNet 5 implicitly uses a priori information about invariants: LeNet 5 includes the module that, along with given examples, considers also examples obtained from the training examples by small randomly picked affine transformations described in a previous section. Using this module, LeNet 5 constructs from one training example 10 new examples belonging to the same class. Therefore it actually uses 600,000 examples. This network outperformed the tangent distance method: It achieved 0.9% error.

The same idea was used in the SV machine. It also used a priori information by constructing virtual examples. The experiment was conducted as follows:

1. Train a SV machine to extract the SV set.
2. Generate artificial examples by translating the support vectors in four main directions (see Fig. 12.7).
3. Train the SV machine again on old SV and generated vectors.

Using this technique, 0.8% performance was obtained. This result was obtained using polynomials of degree 9. Thus in both cases the improvement was obtained due to usage of some a priori information.

12.4.3 The Best Results for NIST Database

The record for NIST database was obtained using the so-called boosting scheme of recognition that combines three LeNet 4 learning machines (Drucker et al., 1993).

The idea of the boosting scheme is as follows. One trains the first learning machine to solve the pattern recognition problem. After completion of training the first machine, one trains the second machine. For this purpose, one uses a new training set from which a subset of training data is extracted,

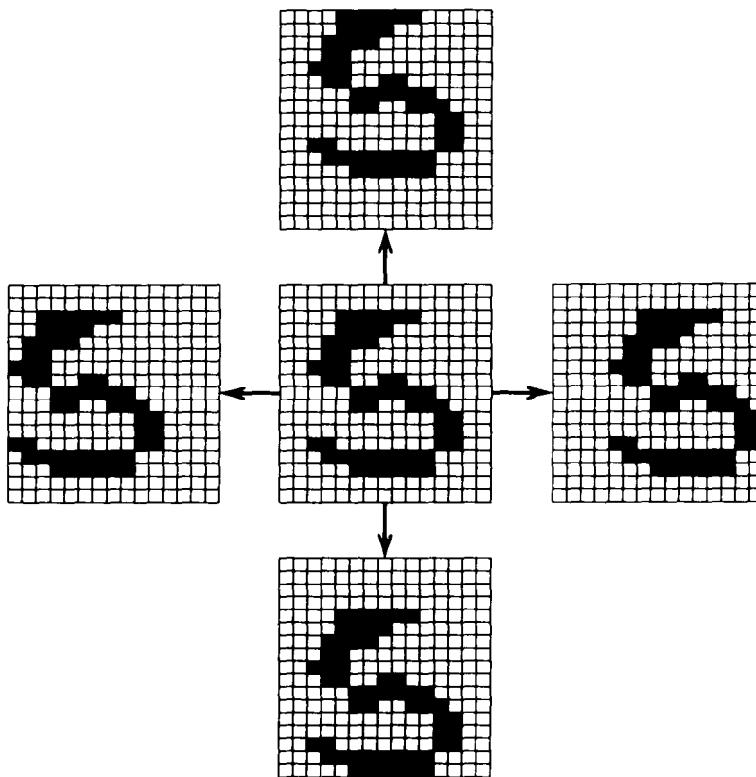


FIGURE 12.7. Image and new examples obtained by translation at two pixels in four main directions (left, right, up, and down).

containing 50% of the examples (chosen randomly) that are correctly classified by the first learning machine and 50% of examples that are incorrectly classified by the first learning machine. After the second machine constructs a decision rule using this subset of training data, a third learning machine is constructed. To do this, a new training set is used from which one chooses examples that the first two machines classify differently. Using this subset of training data, one trains the third machine. Therefore one constructs three different decision rules. The idea of boosting scheme is to use a combination of these three rules for classification (for example, using the majority vote).

It is clear that to use this scheme, one needs a huge amount of training data. Recall that LeNet 4 makes only 1.1% of training error. To construct a second machine, one needs a subset that contains 10,000 errors of the first classifier. This means that one needs a huge amount of new examples to create the second training set. Even more examples are needed to create a training set for the third machine. Therefore in a pure way this scheme looks unrealistic.

Drucker et al. (1993) suggested using a boosting scheme to incorporate a priori knowledge about invariants of handwritten digits with respect to small transformations defined by (12.20). They suggested first to train the learning machine using 60,000 training examples from the NIST database. Then, to get a "new" set of examples they suggested the following "random"

generator of handwritten digits. From any pair (x_i, y_i) of initial training data, they construct new data that contain the same y_i and the new vector

$$x_i(\text{new}) = x_i + \sum_{r=1}^7 a_i x_i^{(r)},$$

where $x^{(r)}$, $r = 1, \dots, 7$ are Lie derivatives described in Section 4.3, $|a_i| \leq C$ is a random vector, and C is reasonably small. In other words, from any given example of initial training data they create new transformed random examples (depending on random vector a_i) that have the same label.

Using three learning machines, LeNet 4 and several million generated examples (obtained on the basis of 60,000 elements of training data), they achieved the performance of 0.7% error rate. Up to now this is the best performance for this database.[†]

12.5 FUTURE RACING

Now the SV machines have a challenge — to cover this gap (between 0.8% to 0.7%). To cover the gap, one has to incorporate more a priori information about the problem at hand.

In our experiments we used only part of available a priori information when we constructed virtual examples by translating support vectors in the four main directions. Now the problem is to find efficient ways of using all available information. Of course using the support vectors, one can construct more virtual examples on the basis of other invariants or one can incorporate this information using a boosting scheme for SV machines.

However, it would be much more interesting to find a way how to incorporate a priori information by choosing appropriate kernels.

The following observations can be used for this purpose.

Observation 1. The best result for the SV machine described in the previous section was obtained using polynomials of degree nine. Recall that these polynomials were constructed in 400-dimensional pixel space that is in $\approx 10^{23}$ -dimensional feature space. Of course most of these coordinates are not useful for constructing the decision rule. Therefore if one could detect these useless terms a priori (before the training data are used), one could reduce the dimensionality of feature space and construct the optimal separating hyperplane in reduced feature space. This will allow us to construct more accurate decision rules.

[†] Note that the best performance 0.8% for the SV machine was obtained using full-size polynomials of degree 9. However, for postal service database using sparse polynomials (to be described in the next section) we significantly improved performance. We hope that the sparse polynomial of degree 9 ($d_1 = 3, d_2 = 3$) will also improve this record.

Let us make the conjecture that high accuracy decision rules are described by the following three-level structures:

1. On the first level a set of local features is constructed. The local features are defined as a product of values of d_1 pixels (say $d_1 = 3$) that are *close to each other*.
2. On the second level a set of global features is constructed. The global features are defined as a product of values of d_2 local features (say $d_2 = 3$).
3. On the third level the optimal hyperplane in the space of global features is constructed.

These decision rules are polynomials of order $d_1 d_2$ (nine in our case) with a reduced number of terms (they are sparse polynomials of degree nine).

The idea of constructing such sparse polynomials can be implemented by constructing a special form of inner product (see Fig. 12.8). Consider patches of pixels, say 4×4 , that overlap over two pixels. Altogether there are 100 such patches. Now let us define the inner product in the form

$$K(u, v) = \left(\sum_{\text{patches}} \left(\sum_{i \in \text{patch}} u_i v_i + 1 \right)^{d_1} \right)^{d_2},$$

where $d_1 = 3$ and $d_2 = 3$. It is easy to see that using this inner product we construct polynomials of degree nine that reflect our conjecture but contain much less terms. The number of terms generated by this inner product is $\approx 10^{14}$ (instead of 10^{23}).

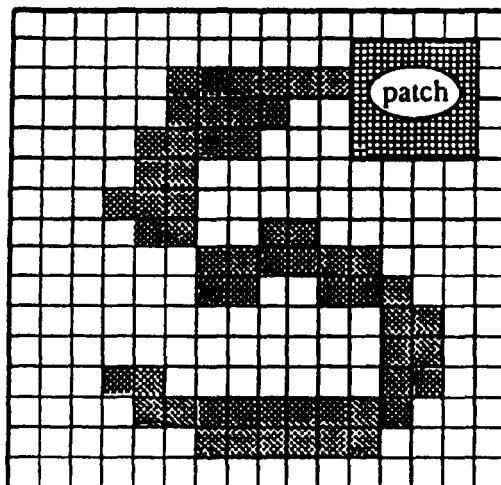


FIGURE 12.8. By constructing special type of inner product, one constructs a set of sparse polynomials

The idea of constructing local features can be used not only for constructing polynomials, but for other kernels as well. Using analogous types of construction of the inner product, one can suggest various structures that reflect a priori knowledge about the problem to be solved.[†]

Observation 2. To construct separating decision rules (say polynomials) that have some invariant properties with respect to transformations described in Section 12.2 (suppose we consider smoothed images), we calculate for any vector x_j of the training data

$$(y_1, x_1), \dots, (y_\ell, x_\ell)$$

the functions $x_j^{(i)}$, $i = 1, \dots, 7$. Consider a high-dimensional feature space Z where the problem of constructing polynomials is reduced to constructing hyperplanes with the inner product defined by the kernel

$$(z_1 * z_2) = K(x_1, x_2).$$

The images of training data in Z space are

$$(y_1, z_1), \dots, (y_\ell, z_\ell).$$

Let $z(x_j)$ be an image in Z space of the vector x_j . Let us define the images of the functions $x_j^{(i)}$, $i = 1, \dots, 7$, as follows:

$$z_j^{(i)} = \lim_{\gamma \rightarrow 0} \frac{z(x_j + \gamma x_j^{(i)}) - z(x_j)}{\gamma}$$

To construct decision rules that are invariant, say, with respect to small horizontal translations (with respect to $x_j^{(1)}$) means to construct the hyperplane

$$(z * \psi) + b = 0$$

such that ψ and b minimize the functional

$$\Phi = \sum_{j=1}^{\ell} \left((z_j^{(1)} * \psi) \right)^2 + C \sum_{j=1}^{\ell} \xi_j \quad (12.22)$$

and satisfy the constraints

$$y_j ((z_j * \psi) + b) \geq 1 - \xi_j,$$

$$\xi_j \geq 0.$$

[†] After this chapter was written, B. Scholkopf reported that by using kernels for sparse polynomials of degree four ($d_1 = 2$, $d_2 = 2$) he obtained 2.9% of the error rate performance on the postal service database.

To obtain the solution, one takes into account that the following equalities are valid:

$$(z_j * \psi) = \sum_{k=1}^{\ell} y_k \alpha_k K(x_j, x_k),$$

$$(z_j^{(1)} * z_i) = \lim_{\gamma \rightarrow 0} \frac{K(x_j + \gamma x_j^{(1)}, x_i) - K(x_j, x_i)}{\gamma} = K^{(1)}(x_j, x_i). \quad (12.23)$$

Here we denote by $K^{(1)}(x_i, x_j)$ the derivative of function $K(x_j, x_i)$ in direction $x_j^{(1)}$. For the polynomial kernel

$$K(x_j, x_i) = (x_j * x_i)^d$$

we have

$$K^{(1)}(x_j, x_i) = d(x_j * x_i)^{d-1} (x_j^{(1)} * x_i).$$

Now, to find the desired decision rule

$$\sum_{i=1}^{\ell} y_i \alpha_i K(x, x_i) + b = 0,$$

one has to solve the following quadratic optimization problem: Minimize the functional

$$\Phi = \sum_{j=1}^{\ell} \left(\sum_{i=1}^{\ell} y_i \alpha_i K^{(1)}(x_j, x_i) \right)^2 + C \sum_{j=1}^{\ell} \xi_j \quad (12.24)$$

subject to constraints

$$y_j \left(\sum_{i=1}^{\ell} y_i \alpha_i K(x_j, x_i) + b \right) \geq 1 - \xi_j, \quad (12.25)$$

$$\sum_{i=1}^{\ell} y_i \alpha_i = 0,$$

$$\alpha_j \geq 0, \quad \xi_j \geq 0, \quad j = 1, \dots, \ell$$

To construct a polynomial that is invariant with respect to several small transformations defined by $x_j^{(i)}$, we construct the decision rule

$$\sum_{i=1}^{\ell} y_i \alpha_i (x * x_i)^d + b = 0$$

such that a , and b minimize the functional

$$\Phi = \sum_{r=1}^{\gamma} \sum_{j=1}^{\ell} \left(\sum_{i=1}^{\ell} y_i \alpha_i d(x_j * x_i)^{d-1} (x_j^{(r)} * x_i) \right) + C \sum_{j=1}^{\ell} \xi_j \quad (12.26)$$

subject to constraints

$$\begin{aligned} y_j \left(\sum_{i=1}^{\ell} y_i \alpha_i (x_j * x_i)^d + b \right) &\geq 1 - \xi_j, \\ \sum_{i=1}^{\ell} y_i \alpha_i &= 0, \\ \alpha_j &\geq 0, \quad \xi_j \geq 0, \quad j = 1, \dots, \ell. \end{aligned}$$

12.5.1 One More Opportunity. The Transductive Inference

There is one more opportunity to improve performance in the digit recognition problem—that is, use transductive inference. Note that the goal of handwritten digit recognition is to read documents that usually contain not one but several digits. For example, read zip codes that contain five digits, or read the courtesy account on bank checks, and so on. The technology of recognition that suggests the framework of the inductive approach is the following: First construct decision rules and then use these rules for recognition of new data.[†]

This idea implies the following sequence of actions: Recognize the first digit (say of a zip code), then the second digit, and so on. Unlike the character recognition in the words you read here, there are no correlations between digits and therefore one recognizes digits independently.

Consider now the same problem in the framework of transductive inference. According to this approach, our goal is to estimate values of the function at the given five points, describing zip codes. We are given training data

$$(y_1, x_1), \dots, (y_\ell, x_\ell) \quad (12.27)$$

(containing thousands of observations) and five new vectors

$$x_1^*, \dots, x_5^*, \quad (12.28)$$

the classification of which

$$y_1^*, \dots, y_5^*$$

is unknown. The goal is to make the classification.

[†] We do not discuss the very difficult problem of digit segmentation (separating one digit from the other for cursive writing digits). In both approaches—the classical one and the new one—we assume that this problem is solved and the main problem is to recognize digits.

Let us assume for a moment that we are faced with a two-class classification problem. Then there exist 32 different ways to classify these five vectors into two classes that contain the desired one. Consider all possible classifications

$$(y_1^1, \dots, y_5^1), \dots, (y_1^{32}, \dots, y_5^{32}).$$

Suppose for simplicity that the collection of the training data and the correctly labeled fives vectors can be separated without error.

Then, according to the theory of estimating values of functions at given points, described in Chapter 8 and Chapter 10 in order to bound the probability of correct classification y_1^k, \dots, y_i^k of five vectors by a linear classifier, one has to estimate the value $D^2(k)/\rho^2(k)$ using the joint set of data containing the training data (12.26) and the data

$$(y_1^k, x_1^*), \dots, (y_5^k, x_5^*)$$

(recall that if classification y_1^k, \dots, y_5^k without error is impossible, then $\rho(k) = 0$). It was shown that the probability of correct classification can be estimated by the value

$$p_{\text{corr}}(k) \geq 1 - \Phi \left(\frac{\min \left(N_{sv}, \frac{D^2(k)}{\rho^2(k)} \right)}{\ell} \right),$$

where $\Phi(u)$ is monotonic. Therefore, to get the best guarantee of classification of data (12.27), one has to choose such a separation (from 32 possible) for which the value

$$d(k) = \min \left(N_{sv}, \frac{D^2(k)}{\rho^2(k)} \right)$$

is the smallest.

Let us now come back to the 10-class classification problem. Up to this point, to conduct a 10-class classification we used 10 two-class classifiers and chose a classification that corresponds to the largest output of the classifier. For estimating the values of a function at given points we also will combine

Table 12.10. Quality of 32 various separations in 10 two-class classification problems

	1	2	32
0	$d_0(1)$	$d_0(2)$	$d_0(32)$
1	$d_1(1)$	$d_1(2)$	$d_1(32)$
...
9	$d_9(1)$	$d_9(2)$	$d_9(32)$

10 two-class classifiers to construct a 10-class classifier. However, the rules for combining these 10 classifiers are more complex.

Consider the following 10 one-class classification problems: digit zero versus the rest, digit one versus the rest, and so on. For any of these 10 problems, one has 32 different possible classifications of our data (12.27). Therefore one can define the table that contains 10 lines, each of which defines the quality of all 32 possible solutions of each two-class classification problem (each column defines the quality of the corresponding solution).

To find the best 10-class classifications, one has to find such 10 two-class solutions that:

1. They are admissible (each element belongs to one class, and there are no contradictions among 10 two-class solutions). There are 100,000 such admissible solutions.
2. Among admissible solutions (sets of 10 two-class solutions), find such for which the score

$$D = \sum_{i=0}^9 d_i(k_i)$$

is minimal, where k_i is the number of chosen solutions in the two-class classification problem for digit i versus the rest.

SV MACHINES FOR FUNCTION APPROXIMATIONS, REGRESSION ESTIMATION AND SIGNAL PROCESSING

This chapter describes examples of solving various real-valued function estimation problems using SV machines. We start with a discussion of the model selection problem and then consider examples of solving the following real-valued function estimation problems:

1. Approximation of real-valued functions defined by collections of data.
2. Estimation of regression functions.
3. Solving the Radon integral equation for positron emission tomography.

13.1 THE MODEL SELECTION PROBLEM

In Chapter 12 when we constructed decision rules for real-life pattern recognition problems we saw the importance of choosing a set of indicator functions with appropriate value of capacity.

For estimating real-valued function the problem of choosing appropriate capacity of an admissible set of functions is even more important.

In Chapter 6 we suggested the principle for choosing such a set of functions —the Structural Risk Minimization (SRM) principle. We suggested that one uses the functional that bounds the risk using information about the empirical risk and the VC dimension of the set of functions of the learning machine.

The main question in the practical application of the SRM principle is the following:

Are the bounds developed in the theory (Chapter 5) accurate enough to be used in practical algorithms for model selection?

In the next two sections we try to answer this question. We describe experiments with model selections which show that for small sample sizes (note that real-life problems are always small sample-size problems) the models selected on the basis on the VC bounds are more accurate than the models selected on the basis of classical statistical recommendations.

13.1.1 Functional for Model Selection Based on the VC Bound

We start with a simple particular problem of model selection. Given a collection of data, estimate a regression function in the set of polynomials, where the order of the best approximating polynomial is unknown and has to be estimated from the data.

Let in the interval $[a,b]$ the probability density $p(x)$ be defined and let there exist a conditional density $p(y|x)$ that defines values of y for a given vector x . Therefore the joint probability distribution function

$$p(x, y) = p(x)p(y|x)$$

is defined. Let

$$(y_1, x_1), \dots, (y_\ell, x_\ell)$$

be i.i.d. data governed by this distribution function. Our goal is to use the data to approximate the regression function

$$r(x) = \int y p(y|x) dy$$

by some polynomial. Note that the regression function is not necessarily a polynomial.

To find the best approximating polynomial one has to answer two questions:

1. What is the best order of approximating polynomial?
2. What are the parameters of this polynomial?

The second question has a simple answer. One chooses the parameter a that minimizes the empirical risk functional (say with quadratic loss function)

$$R_{\text{emp}}(\alpha) = \frac{1}{\ell} \sum_{i=1}^{\ell} \left(y_i - \sum_{k=0}^d \alpha_k x_i^k \right)^2.$$

One cannot, however, use this functional to choose the appropriate order d of the approximating polynomial, since the value of this functional decreases with increasing order of polynomial and becomes equal to zero when $d =$

$\ell - 1$. Therefore to choose the appropriate order of polynomial, one has to minimize another functional.

Classical statistics suggest several solutions to this problem. Each of these suggests some functional to be minimized instead of the empirical risk functional. In the next section we will describe these functionals. Below we consider the problem of choosing the order of polynomial as a problem of choosing the appropriate element of the structure in the scheme of structural risk minimization.

Note that the setting of this problem actually describes the structure on the set of polynomials. Indeed, in the problem of finding the best order of polynomial, the first element S_1 of the structure is the set of constant functions, the second element S_2 is the set of linear functions, and so on. For any element of this structure we have a good estimate of the VC dimension. Since the set of polynomials of order $k - 1$ is a set of functions linear in k parameters, the VC dimension of element S_k is equal to k (the number of free parameters). As we saw in Chapters 4 and 5, the number of free parameters is a good estimate of the VC dimension for a linear set of functions. According to the bounds obtained in Chapter 5, if the distribution function is such that the inequality

$$\sup_k \sup_{\alpha \in R^k} \frac{\sqrt[p]{E(y - P_k(x, \alpha))^2 p}}{E(y - P_k(x, \alpha))^2} \leq \tau < \infty, \quad p > 1$$

is valid, where $P_k(x, \alpha)$ denotes the polynomial of degree $k - 1$, then with probability $1 - \eta$ simultaneously for all α the inequality

$$E(y - P_k(x, \alpha))^2 \leq \left(\frac{\frac{1}{\ell} \sum_{i=1}^{\ell} (y_i - P_k(x_i, \alpha))^2}{1 - c(\tau, p) \sqrt{\frac{k \left(\ln \frac{2\ell}{k} + 1 \right) - \ln \frac{\eta}{4}}{\ell}}} \right)_+ \quad (13.1)$$

holds, where $c(\tau, p)$ is a constant depending only on τ and p . According to the SRM principle to find the best solution one has to minimize the right-hand side of (13.1) with respect to the parameter k (the order of the polynomial) and parameters α (coefficients of polynomials). To solve the problem of choosing the appropriate order of polynomial, we use the functional that differs from (13.1) only by constants. We also will specify the choice of η depending on ℓ as follows:

$$\eta = \frac{4}{\sqrt{\ell}}.$$

Thus to choose the order of polynomial we minimize the functional (Vapnik,

1979)

$$R(k, \alpha) = \frac{\frac{1}{\ell} \sum_{i=1}^{\ell} (y_i - P_k(x_i, \alpha))^2}{\left(1 - \sqrt{\frac{k(\ln \frac{\ell}{k} + 1) + \frac{\ln \ell}{2}}{\ell}} \right)_+}. \quad (13.2)$$

13.1.2 Classical Functionals

In the previous section we defined the functional that has to be minimized instead of empirical risk functional. This functional has the form

$$R(k, \alpha) = g(k, \ell) \frac{1}{\ell} \sum_{i=1}^{\ell} (y_i - P_k(x_i, \alpha))^2. \quad (13.3)$$

The recommendations that come from classical analysis also suggest minimization of the functional of type (13.3) with different correcting functions $g(k, \ell)$. The following describes the four most popular of these recommendations.

Finite prediction error (FPE) (Akaike, 1970) uses the following correcting function

$$g(k, \ell) = \frac{1 + \frac{k}{\ell}}{1 - \frac{k}{\ell}}. \quad (13.4)$$

Generalized cross-validation (GCV) (Craven and Wahba, 1979) uses the following correcting function

$$g(k, \ell) = \frac{1}{\left(1 - \frac{k}{\ell} \right)^2}. \quad (13.5)$$

Shibata's model selector (SMS) (Shibata, 1981) uses the following correcting function:

$$g(k, \ell) = 1 + 2 \frac{k}{\ell}. \quad (13.6)$$

Schwartz criteria (MDL criteria) (Schwartz, 1978) uses the following correcting function:

$$g(k, \ell) = 1 + \frac{\frac{k}{\ell} \ln \ell}{2 \left(1 - \frac{k}{\ell} \right)}. \quad (13.7)$$

Note that first three correcting functions have the same asymptotic form

$$g(k, \ell) = 1 + 2\frac{k}{\ell} + O\left(\frac{k^2}{\ell^2}\right)$$

In the next section we compare the performance obtained using these classical correcting functions with the functional (13.2) obtained on the basis of the VC bound.

13.1.3 Experimental Comparison of Model Selection Methods

The following experiments were conducted for each model selection functional (Cherkassky et al., 1996). Using a set of polynomials we tried to approximate the nonpolynomial regression function $\sin^2 2\pi x$ corrupted by noise:

$$y = \sin^2 2\pi x + \varepsilon.$$

We considered the training data

$$(y_1, x_1), \dots, (y_\ell, x_\ell)$$

obtained on the basis of a uniform distribution x on the interval $[0,1]$ and normally distributed noise with zero mean and different values of variance. The model selection criteria were used to determine the best approximation for a given size of training data. The mean-squared deviation of the chosen approximation function from the true function was used to evaluate performance.

Four different sizes of training data (10, 20, 30, 100 samples) with 10 different levels of noise were tried. The noise is defined in terms of signal-to-noise ratio (SNR) given by the mean-squared deviation of the signal from its mean value to the variance of the noise.

All experiments were repeated 1000 times for a given training set size and noise level. Therefore for any experiment we could construct a distribution function on performances. Schematically we describe these distribution functions using standard box notation (see Fig. 13.1).

Standard box notation specifies marks at 95, 75, 50, 25, and 5 percentile of an empirical distribution. The results of these experiments are presented in Figs. 13.2, 13.3, and 13.4. These figures show the distribution of mean-squared deviation of the approximating function from regression obtained for five functionals (FPE, GCV, SMS, MDL, VC) of the model selection, different signal/noise ratio (SNR) and different number n of observations.

Along with these experiments, similar experiments for other target functions were conducted. They showed similar results in performance of various methods.

From these experiments, one can conclude the following:

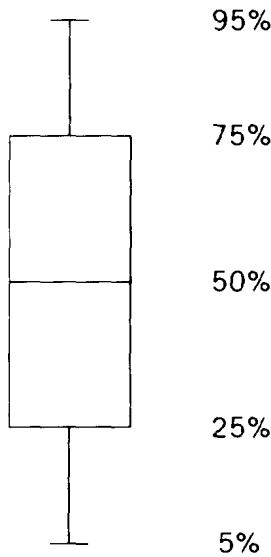


FIGURE 13.1. Standard box notation for description of results of statistical experiments. It specifies marks at 95, 75, 50, 25, and 5 percent of an empirical distribution.

1. For small sample size, classical methods did not perform particularly well.
2. For small sample size the functional (13.2) gives consistently reasonable results over the range tested (small error as well as small spread).
3. Performance for large samples (more than 100; they are not presented on the figures) is approximately the same for all methods for the amount of noise used in this study.

13.1.4 The Problem of Feature Selection Has No General Solution

The generalization of the problem of choosing an order of the approximating polynomial is the following: Given an ordered sequence of features and the structure where the element S_k contains the functions constructed as a linear combination of the first k features. It is required to choose the best element of the structure to estimate the desired function using the data. This problem can be solved using functional (13.2)

In practice, however, it is not easy to order a collection of admissible features *a priori*. Therefore one tries to determine the order on the set of features, the appropriate number of ordered features, and the decision rule using training data. In other words, using the data one would like to choose among a given set of features $\{\psi(x)\}$ a small number of appropriate features, say $\psi_1(x), \dots, \psi_n(x)$, and then using the same data construct a model

$$y = \sum_{k=1}^n \alpha_k \psi_k(x). \quad (13.8)$$

However, to select an appropriate number n in this case, one cannot use functional (13.2) for the following reason: In contrast to the problem of choosing

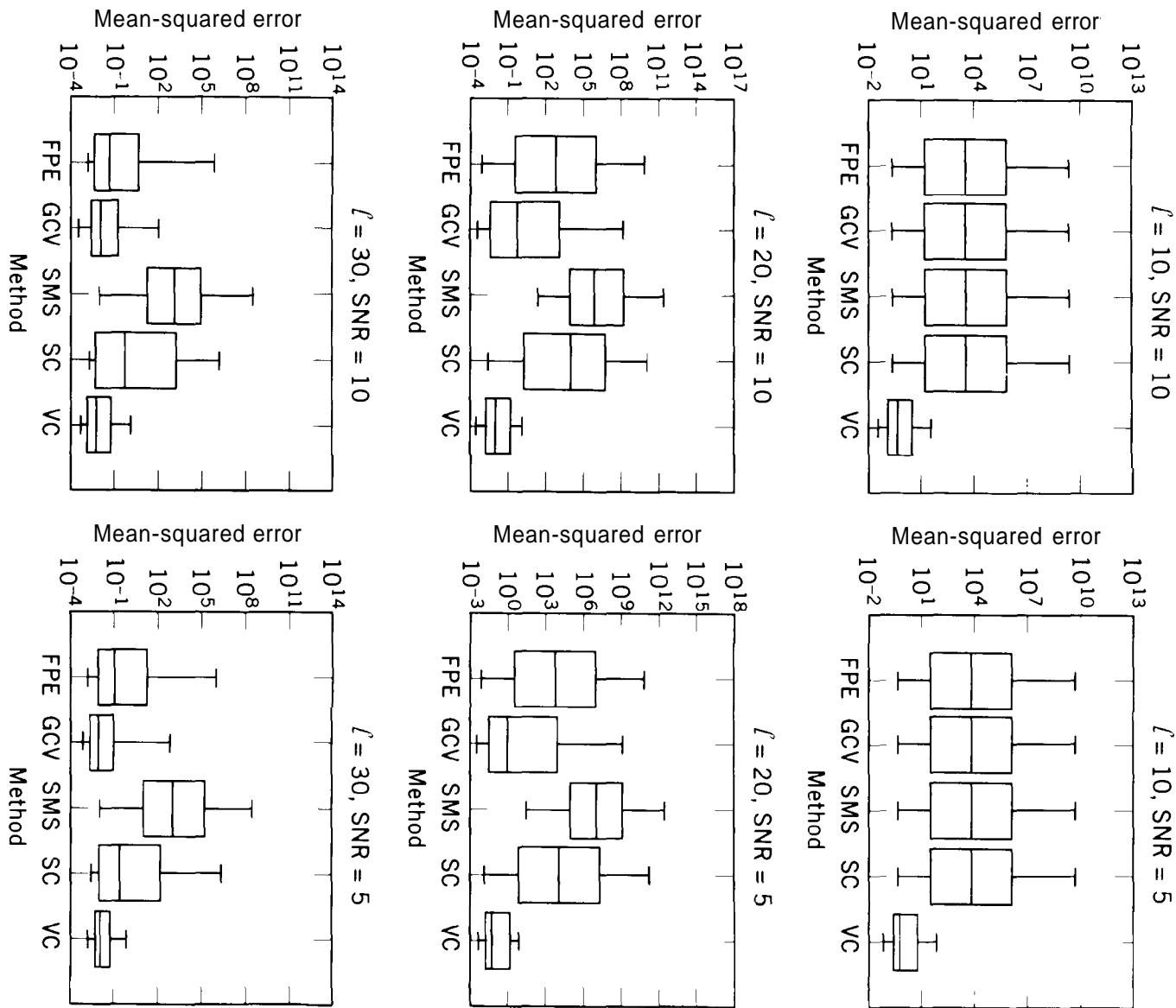


FIGURE 13.2. Results of experiments for different methods, different number of observations, and different values of SNR.

the order of approximating polynomial in this scheme of structural risk minimization, one cannot specify a priori which features will be chosen first and which will be chosen next. A priori all combinations are possible. Therefore the n th element of structure contains functions that are linear combination of any n features.

It turns out that the characteristics of the capacity for elements of such a structure depend not only on the number of terms n in the linear model (13.8), but also on the set $\{\psi(x)\}$ from which the features were chosen.

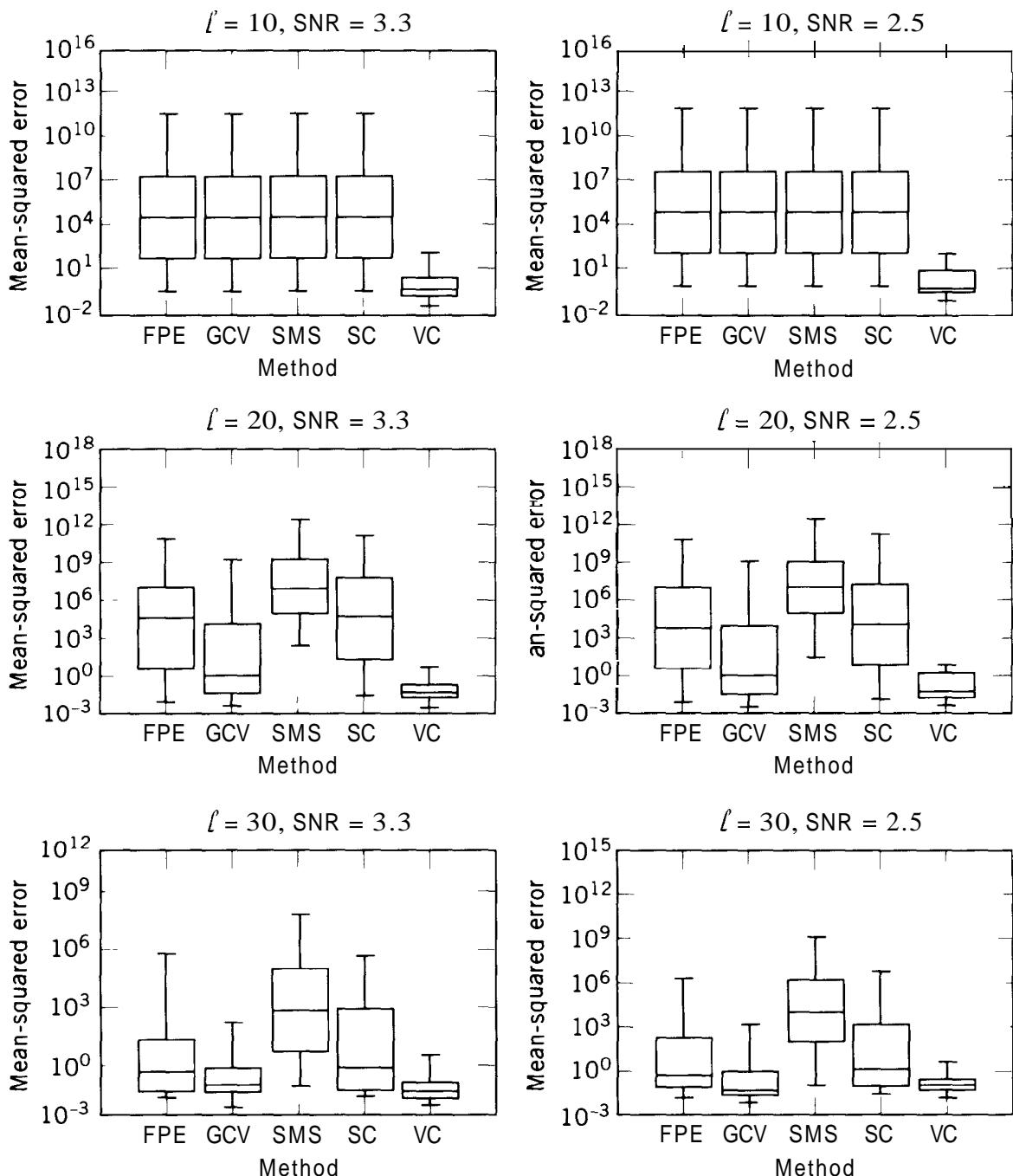


FIGURE 13.3. Results of experiments for different methods, different number of observations, and different values of SNR.

Consider the following situation. Let our set of one-dimensional functions be polynomials. That is,

$$\psi_0(x) = 1, \quad \psi_1(x) = x, \dots, \quad \psi_n(x) = x^n, \dots$$

Suppose now that we would like to select the best n features to construct the

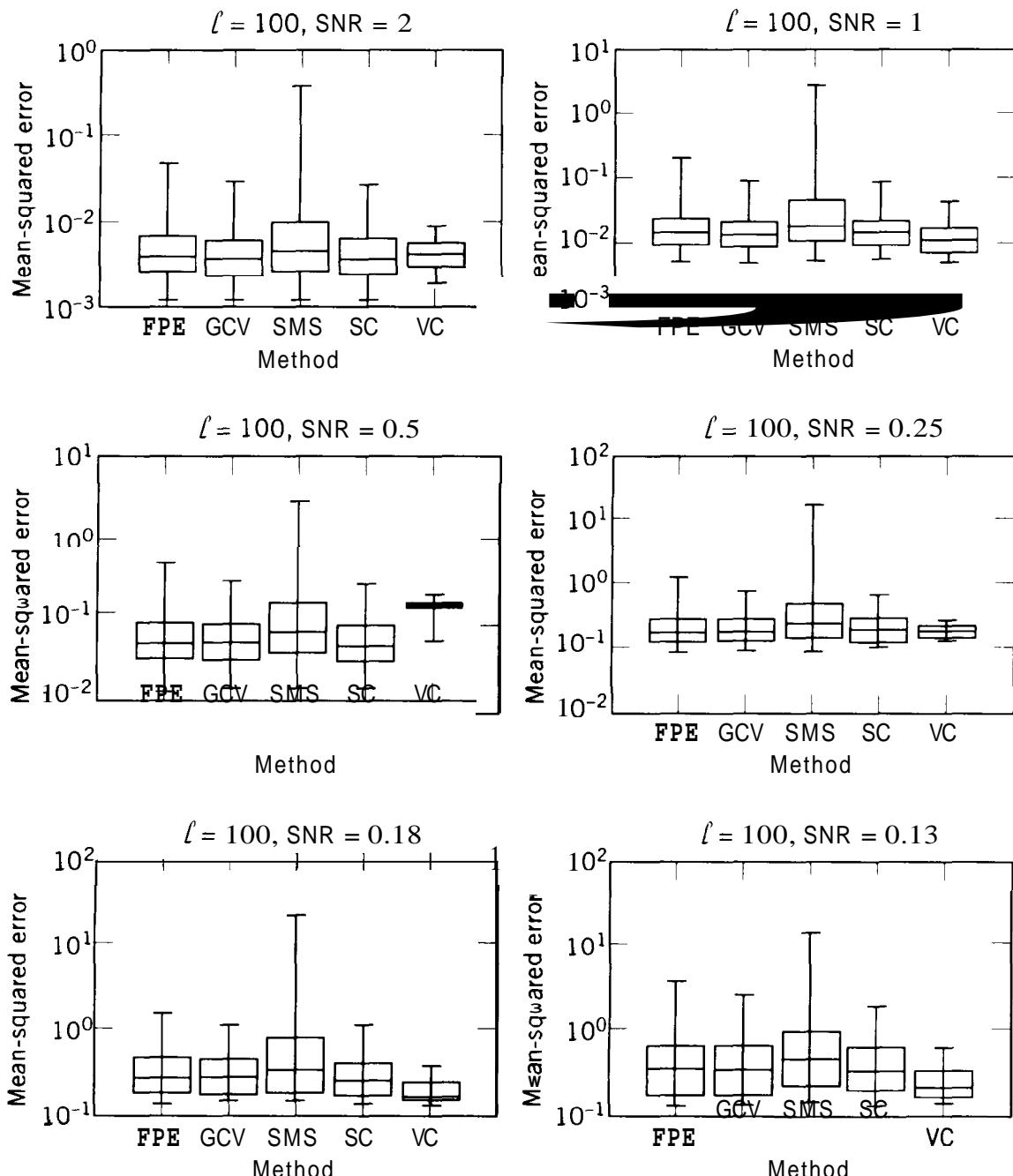


FIGURE 13.4. Results of experiments for different methods, different number of observations, and different values of SNR.

model

$$y = \sum_{k=1}^n \alpha_k \psi_{n_k}(x), \quad 0 \leq x \leq 1.$$

This function can be a polynomial of any order, but it contains only n terms—the so-called sparse polynomials with n terms (compare with the problem of choosing an order of approximating polynomial where element S_n of the structure contained all polynomials of degree $n - 1$).

It is known (Karpinski and Werther, 1989) that the VC dimension for a set of sparse polynomials with n terms has the following bounds:

$$3n - 1 \leq h_n \leq 4n. \quad (13.9)$$

Therefore to choose the best sparse polynomial, one can use the functionals defined (13.2), where instead of $h_n = n$ one has to use $h_n = 4n$.

Now let us consider another set of features

$$\psi_0(x) = 1, \quad \psi_1(x) = \sin \pi x, \dots, \quad \psi_n(x) = \sin n\pi x, \dots, \quad 0 \leq x \leq 1$$

from which one has to select the best n features in order to approximate a desired function. For this set of features the problem has no solution, since as we showed in Chapter 4, Section 4.11 (Example 4) the VC dimension of the set of functions $\sin ax$ is infinite, and therefore it can happen that by using one feature from this set (with sufficiently large a), one can approximate the data but not the desired function.

Thus, the problem of feature selection has no general solution. The solution depends on the set of admissible features from which one chooses the appropriate n ones.

One can consider a particular case where the number of features from which one chooses an appropriate feature is finite, say equal to N . In this case the bound on capacity of element S_n of the structure is $h_n \leq n \ln N$. This bound, however, can be both relatively accurate for one set of functions (say, for trigonometric polynomials) and pessimistic for another set of functions (say, for sparse algebraic polynomials, where $h_n \leq 4n$). Therefore using only information about the number of features used to obtain a specific value of empirical risk and the number of admissible features one cannot control the generalization well.

13.2 STRUCTURE ON THE SET OF REGULARIZED LINEAR FUNCTIONS

Consider another idea of constructing a structure on the set of functions linear in their parameters that was suggested in the early 1960s for solving ill-posed problems.

Consider a set of functions linear in their parameters

$$y = (\mathbf{a} * \mathbf{x})$$

and the positive functional

$$\Omega(\mathbf{a}) = (\mathbf{a} * \mathbf{A}\mathbf{a}).$$

Suppose our goal is given the data

$$(x_1, y_1), \dots, (x_\ell, y_\ell)$$

to estimate the regression function. Consider the following a priori structure on the set of linear functions

$$S_k = \{a : (a * Aa) \leq c_k\}, \quad (13.10)$$

where A is a positive definite matrix and c_k , $k = 1, \dots$, is a sequence of monotonic increasing constants.

To find the regression function we minimize the functional

$$R(a) = \frac{1}{\ell} \sum_{i=1}^{\ell} (y_i - (a * x_i))^2 \quad (13.11)$$

on the element S_k of the structure (13.10). This problem is equivalent to minimizing the functional

$$R(a) = \frac{1}{\ell} \sum_{i=1}^{\ell} (y_i - (a * x_i))^2 + \gamma_k (a * Aa), \quad (13.12)$$

where the choice of the nonnegative constant γ_k is equivalent to the choice of elements of the structure (13.10) (choice of constant c_k in Eq. (13.10)).

For SV machines (without loss of generality (see Section 10.9)) we consider a particular case of this structure—namely, the case where A in (13.10) is the identity matrix $A = I$. In this case the functional (13.12) has a form

$$R(a) = \frac{1}{\ell} \sum_{i=1}^{\ell} (y_i - (a * x_i))^2 + \gamma_k (a * a). \quad (13.13)$$

Below we discuss three heuristic methods for choosing the regularization parameter γ_k which came from different theories that consider (from a different point of view) the regularization problem:

1. The L-curve method, which was suggested for solving ill-posed problem. (Hansen, 1992; Engl et al., 1996).
2. The effective number of parameters method, which was suggested in statistics for estimating parameters of ridge regression (in statistical literature the minimum of functional (13.13) is called ridge-regression) (Hestini and Tibshirani, 1990; Wahba, 1990; Moody, 1992).
3. The effective VC dimension, which was developed in the framework of statistical learning theory (Vapnik, Levin, and LeCun, 1994).

13.2.1 The L-Curve Method

Suppose that for a fixed y the vector a_γ provides the minimum to the functional (13.13). Consider two functions:

$$s(\gamma) = \ln \left[\frac{1}{\ell} \sum_{i=1}^{\ell} (y_i - (a_\gamma * x_i))^2 \right]$$

and

$$t(\gamma) = \ln [(a_\gamma * a_\gamma)].$$

One can consider these two functions as a parameterized form (with respect to parameter γ) of the function

$$t = L(s)$$

This curve looks like a character "L" (that is why this method was called the L-curve method): For large γ the value of t is small but the value of s is large, and with decreasing γ the value of t increases but the value of s decreases. It was observed that a good value of the regularization parameter γ corresponds to the corner point of the curve.

Let us call point $\mathcal{L} = (t(\gamma^*), s(\gamma^*))$ the corner point of curve $t = L(s)$ if it satisfies two properties:

1. The tangent of $t = L(s)$ at $s^* = s(\gamma^*)$ has a slope equal to -1 :

$$\frac{dL(s^*)}{ds} = -1$$

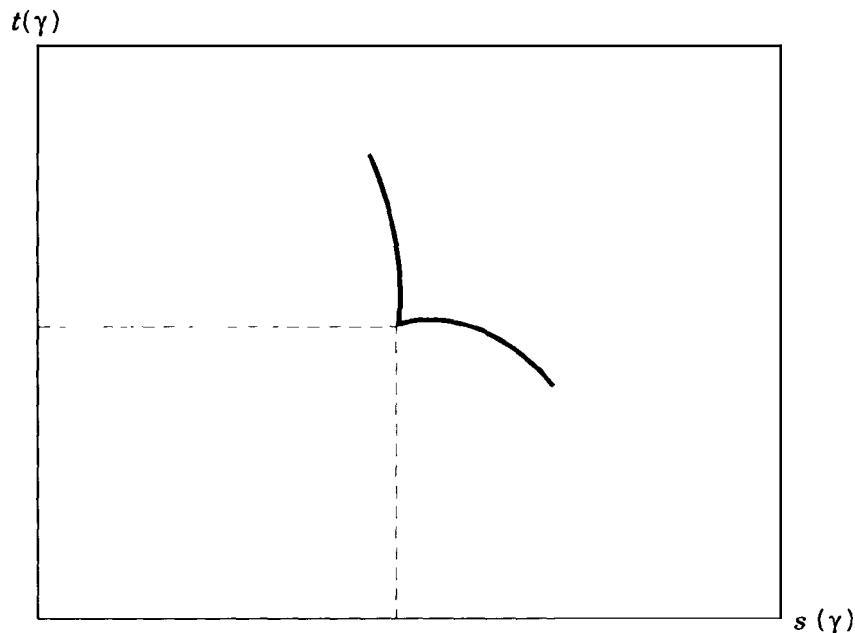


FIGURE 13.5. The form of L curve. The parameter γ that corresponds to the corner point on the curve defines the value of regularization.

2. The function $L(s)$ is concave in the neighborhood of $s(\gamma^*)$.

The value y that defines this point has to be chosen for regularization.

It is easy to show that under the condition that $t(\gamma)$ and $s(\gamma)$ are differentiable functions, the point $(t(\gamma^*), s(\gamma^*))$ is the corner point if and only if the function

$$\mathcal{H}(\gamma) = \frac{1}{\ell} \sum_{i=1}^{\ell} (y_i - (a_{\gamma} * x_i))^2 (a_{\gamma} * A a_{\gamma})$$

has a local minima at $y = y^*$.

Indeed, since the function $\mathcal{H}(\gamma)$ can be written as

$$\mathcal{H}(\gamma) = \exp\{t(\gamma) + s(\gamma)\}$$

the necessary condition for its local minimum is

$$\frac{d\mathcal{H}(\gamma^*)}{d\gamma} = \left(\frac{dt(\gamma^*)}{d\gamma} + \frac{ds(\gamma^*)}{d\gamma} \right) \exp\{t(\gamma) + s(\gamma)\} = 0.$$

This implies

$$\frac{dt(\gamma^*)}{d\gamma} + \frac{ds(\gamma^*)}{d\gamma} = 0;$$

that is, the tangent of the L curve is -1 . Since y^* is also the local minimum of $\ln \mathcal{H}(\gamma)$ we have

$$t(\gamma) + s(\gamma) > t(\gamma^*) + s(\gamma^*).$$

Thus, the point of local minima \mathcal{H} is the corner point. Analogously, one can check that y^* that defines the corner point provides the local minimum for function $\mathcal{H}(\gamma)$.

The L-curve method can be applied to the SV machine for choosing the regularization constant:

$$C = \frac{1}{\gamma}.$$

The objective function in the feature space for the support vector method with quadratic loss function has form (13.13), and therefore all the above reasoning is valid for SV machines. To find the best parameter $C = 1/\gamma$ for SV machines using the L-curve approach, one has to find a local minimum of the functional

$$\mathcal{H}(\gamma) = \frac{1}{\ell} \sum_{i=1}^{\ell} \left(y_i - \sum_{k=1}^N \beta_k(\gamma) K(x_k, x_i) \right)^2 \sum_{i,j=1}^N \beta_i(\gamma) \beta_j(\gamma) K(x_i, x_j).$$

13.2.2 The Method of Effective Number of Parameters

The statistical approach to the problem of choosing the best regularization parameter is based on the idea to characterize the capacity of the set of regularization functions using the generalized concept of the number of free parameters—the so-called effective number of parameters. The effective number of parameters is used instead of the number of parameters in functionals that estimate the accuracy of prediction.

Let us define the concept of the effective number of parameters. Suppose we are given the training data

$$(y_1, x_1), \dots, (y_\ell, x_\ell).$$

Consider the matrix

$$\mathbf{B} = \mathbf{X}^T \mathbf{X},$$

where \mathbf{X} is $\ell \times n$ matrix of vectors x_i .

Let

$$\lambda_1, \dots, \lambda_n$$

be (nonnegative) eigenvalues of the matrix \mathbf{B} ordered according to decreasing value and let

$$\psi_1, \dots, \psi_n \tag{13.14}$$

be the corresponding eigenvectors of this matrix. The value

$$h_{\text{eff}} = \sum_{i=1}^n \frac{\lambda_i}{\lambda_i + \gamma},$$

which is the trace of the matrix

$$h_{\text{eff}} = \text{trace} \left(\mathbf{X} (\mathbf{B} + \gamma I)^{-1} \mathbf{X}^T \right)$$

is called the effective number of parameters.

The idea of using the effective number of parameters in functionals (13.4), (13.5), (13.6), (13.7), and (13.2) instead of the number of parameters is justified by the following observation. Suppose the vector coefficient a_0 that minimizes functional (13.11) has the following expansion on the eigenvectors

$$a_0 = \sum_{i=1}^n \alpha_i^0 \psi_i$$

and therefore the function that is defined by this vector is

$$y = \sum_{i=1}^n \alpha_i^0 (\psi_i^* x).$$

It is easy to check that the function that is defined by the vector that minimizes the functional (13.13) is

$$y = \sum_{i=1}^n \frac{\lambda_i}{\lambda_i + \gamma} \alpha_i^0 (\psi_i * x). \quad (13.15)$$

Suppose that the (nonnegative) eigenvalues λ_i rapidly decrease with increasing i . Then the values

$$\delta_i = \frac{\lambda_i}{\lambda_i + \gamma}$$

are either close to zero (if $\lambda_i \ll y$) or close to one (if $\lambda_i \gg y$). In this situation by choosing different values of y one can control the number $h_{\text{eff}} = h(\gamma)$ of terms in expansion (13.15) that are essentially different from zero. In other words, if the eigenvalues decrease rapidly, the regularization method realizes the structural risk minimization principle where the structure is defined by expansion on first k eigenvectors. The number k of the element of the structure S_k is defined by the constant y .

The method of effective number of parameters can be implemented for support vector machines that use a reasonable number of training data (say, up to several thousand). Let $K(x_i, x_j)$ be the kernel that defines the inner product in the feature space Z and let

$$(y_1, z_1), \dots, (y_\ell, z_\ell)$$

be training data in the feature space. Let

$$B^* = Z^T Z \quad (13.16)$$

be an $N \times N$ covariance matrix of the training vectors in the feature space.

It is easy to show that the nonzero eigenvalues of B^* matrix coincide with eigenvalues of the $\ell \times \ell$ matrix K defined by the elements $k_{i,j} = K(x_i, x_j)$, $i, j = 1, \dots, \ell$:

$$K = ||K(x_i, x_j)||.$$

Indeed consider the eigenvector corresponding to the largest eigenvalue λ as the expansion

$$V = \sum_{i=1}^{\ell} b_i z_i, \quad (13.17)$$

where b_1, \dots, b_ℓ are coefficients that define the expansion.

According to the definition of eigenvectors and eigenvalues the equality

$$B^* V = \lambda V \quad (13.18)$$

holds true. Putting (13.16) and (13.17) into (13.18), one obtains the equality

$$\sum_{i,j=1}^{\ell} z_i k_{ij} b_j = \lambda \sum_{i=1}^{\ell} b_i z_i$$

from which, multiplying both sides by z_t , one obtains the equality

$$\sum_{i,j=1}^{\ell} k_{ti} k_{ij} b_j = \lambda \sum_{j=1}^{\ell} k_{tj} b_j.$$

The last equality can be rewritten in the short form

$$K K b = \lambda K b,$$

where we denote by b the vector of expansion coefficients b_1, \dots, b_ℓ . Denoting

$$W = Kb$$

we obtain our assertion

$$K W = A W.$$

Therefore for a reasonable number of observations, one can estimate nonzero eigenvalues in the feature space using the standard technique.

Knowing the eigenvalues $\lambda_i, i = 1, \dots, \ell$, one can calculate the effective number of parameters h_{eff} for the SV machine with a quadratic loss function where $y = 1/C$. The effective number of parameters is used in the objective functionals described in Sections 13.1.1 and 13.1.2.

13.2.3 The Method of Effective VC Dimension

Consider a method for estimating the VC dimension of learning machines by measuring it in experiments with the machine itself. The estimated value (let us call it effective VC dimension) can be used in functional (13.2) instead of the actual value of the VC dimension in order to select the appropriate model in the problem of function estimation.[†]

Since the VC dimension of the set of real-valued functions $(x * a) + b$, $a \in A, b \in R^1$ coincides with the VC dimension of the set of indicator functions $\theta\{(x * a) + b\}$, $a \in A, b \in R^1$, it is sufficient to find a method for measuring the VC dimension of the set of indicator functions.

[†]We call the estimated value the effective VC dimension or for simplicity the estimate of VC dimension. However, this value is obtained by taking into account values x of training data and therefore can describe some capacity concept that is between the annealed entropy and the VC dimension. This, however, is not very important since all bounds on generalization ability derived in Chapters 4 and 5 are valid for any capacity concept that is larger than the annealed entropy.

The idea of measuring the VC dimension of the set of indicator functions is inspired by the technique of obtaining the bounds described in Chapter 4. Let

$$(y_1, x_1), \dots, (y_{2\ell}, x_{2\ell}), \quad y \in \{0, 1\}, \quad x \in R^n \quad (13.19)$$

be training data. If in the training data y are real values, create artificial training data with random zero-one values for y .

Using the same technique that was used in Chapter 4 for obtaining the bound on maximal deviation between frequency on two subsamples, one can prove that there exists such a monotonic decreasing-to-zero function $\Phi(t)$, where $t = \ell/h^*$ is the ratio of the number of elements of the data to the capacity, that the following inequality holds

$$\begin{aligned} E \left\{ \sup_{\alpha, b} \left(\frac{1}{\ell} \sum_{i=1}^{\ell} |y_i - \theta\{(x_i * \alpha) + b\}| - \frac{1}{\ell} \sum_{i=\ell+1}^{2\ell} |y_i - \theta\{(x_i * \alpha) + b\}| \right) \right\} \\ \leq \Phi\left(\frac{\ell}{h^*}\right) \end{aligned} \quad (13.20)$$

Using different concepts of capacity one obtains different expressions of the function $\Phi\left(\frac{\ell}{h^*}\right)$. In particular

$$\Phi\left(\frac{\ell}{h^*}\right) \leq \Phi^*\left(\frac{H_{\text{ann}}^\Lambda(2\ell)}{\ell}\right) \leq \Phi^*\left(\frac{G^\Lambda(2\ell)}{\ell}\right) \leq \Phi^*\left(\frac{\ln 2\ell/h + 1}{\ell/h}\right),$$

where $H_{\text{ann}}^\Lambda(2\ell)$ is the annealed entropy, $G^\Lambda(2\ell)$ is the growth function, and h is the VC dimension.

Let us hypothesize that there exists such a capacity parameter h^* called the effective VC dimension (which can depend on an unknown probability measure on X) and such universal function $\Phi\left(\frac{\ell}{h^*}\right)$ that for any fixed set of functions $f(x, a)$ and for any fixed probability measure on X the equality (not only the inequality as (13.20)) is valid:

$$\begin{aligned} E \left\{ \sup_{\alpha, b} \left(\frac{1}{\ell} \sum_{i=1}^{\ell} |y_i - \theta\{(x_i * \alpha) + b\}| - \frac{1}{\ell} \sum_{i=\ell+1}^{2\ell} |y_i - \theta\{(x_i * \alpha) + b\}| \right) \right\} \\ = \Phi\left(\frac{\ell}{h^*}\right) \end{aligned} \quad (13.21)$$

Suppose that this hypothesis is true: There exists a universal function $\Phi(t)$, $t = \ell/h^*$, such that (13.21) holds. Then in order to construct the method for measuring the VC effective dimension of the set of indicator functions $\delta\{(x * a) + b\}$, one needs to solve the following two problems:

1. Using training data define the experiment for constructing random values

$$\xi_{\ell_k} = \sup_{\alpha, b} \left(\frac{1}{\ell_k} \sum_{i=1}^{\ell_k} |y_i - \theta\{(x_i * \alpha) + b\}| - \frac{1}{\ell_k} \sum_{i=\ell_k+1}^{2\ell_k} |y_i - \theta\{(x_i * \alpha) + b\}| \right)$$

2. Find a good approximation to the universal function $\Phi\left(\frac{\ell}{h}\right)$.

Having solved these two problems, estimate the VC dimension as follows:

1. For different ℓ_k define the values ξ_k , that is, define the pairs

$$(\ell_1, \xi_{\ell_1}), \dots, (\ell_k, \xi_{\ell_k}). \quad (13.22)$$

2. Choose the parameter h^* of the universal function $E\xi_\ell = \Phi\left(\frac{\ell}{h^*}\right)$ that provides the best fit to the data (13.22).

How to Construct a Set of Examples. To construct a set of examples (13.22) using the data (13.19) do the following:

1. Choose (randomly) from the set of data (13.19) a subset of size $2\ell_k$
2. Split (randomly) this subset into two equal subsets:

$$(y_1, x_1), \dots, (y_{\ell_k}, x_{\ell_k}) \quad \text{and} \quad (y_{\ell_k+1}, x_{\ell_k+1}), \dots, (y_{2\ell_k}, x_{2\ell_k}).$$

3. Change the labels y , in the first subset to the opposite $\bar{y}_i = 1 - y_i$,
4. Construct the new set of data

$$(y_1^*, x_1), \dots, (y_{\ell_k}^*, x_{\ell_k}), (y_{\ell_k+1}^*, x_{\ell_k+1}), \dots, (y_{2\ell_k}^*, x_{2\ell_k}) \quad (13.23)$$

containing first the subset with changed labels and second the subset with unchanged labels.

5. Minimize the empirical risk

$$R_{\text{emp}}(\alpha, b) = \frac{1}{2\ell_k} \sum_{i=1}^{2\ell_k} |y_i^* - \theta\{(x_i * \alpha) + b\}| \quad (13.24)$$

in a set of indicator functions with parameters $a \in A$. Let α^* and b^* be the parameters that minimize the empirical risk (13.24). Then (ℓ_k, ξ_{ℓ_k}) , where

$$\xi_{\ell_k} = \frac{1}{\ell_k} \sum_{i=1}^{\ell_k} |y_i - \theta\{(x_i * \alpha^*) + b^*\}| - \frac{1}{\ell_k} \sum_{i=\ell_k+1}^{2\ell_k} |y_i - \theta\{(x_i * \alpha^*) + b^*\}|,$$

is the desired pair. Indeed, it is easy to check that

$$\xi_{\ell_k} = 1 - 2R_{\text{emp}}(\alpha^*, b^*). \quad (13.25)$$

6. Repeating this experiment several times for different k , one obtains set (13.22).

How to Approximate the Function $\Phi\left(\frac{\ell}{h}\right)$. To construct the function $\Phi\left(\frac{\ell}{h}\right)$, one uses a machine with known VC dimension h to create the examples (13.22). Then by using these examples one can construct the function $\Phi(t)$, $t = \frac{\ell}{h}$ as a regression.

The idea behind this method of measuring the VC dimension is very simple. Create a sample (13.23) that does not lead to generalization (the probability of correct answers for the problem is 0.5). Estimate the expectations of the empirical risk for different number of such examples. For these examples the expectations of empirical risk can be less than 0.5 only due to the capacity of the set of function but not due to the generalization. Knowing how the expectation of the minimum of empirical risk increases with increasing the number of observations one can estimate the capacity.

Choosing a Regularization Parameter for Ridge Regression. Suppose that our machine can effectively control the capacity. For example the ridge-regression machine that in order to minimize the error, minimizes the functional

$$R(\alpha) = \frac{1}{2\ell_k} \sum_{i=1}^{2\ell_k} (y - (x_i * \alpha) - b)^2 + \gamma(a * Aa). \quad (13.26)$$

For this machine the capacity is controlled by parameter y . For fixed capacity (fixed $y \geq 0$) using various samples of size ℓ_k we obtain the parameters $\alpha_{\ell_k}^*$, $b_{\ell_k}^*$ that provide minimum to functional (13.26) on data (13.23). We use these parameters in (13.24) to estimate $R_{\text{emp}}(\alpha^*, b^*)$ and then use the values ξ_{ℓ_k} (13.25) in sequence (13.22).

To estimate the effective VC dimension we use the following universal function $\Phi(t)$, $t = \left(\frac{\ell}{h}\right)$:

$$\Phi(t) = \begin{cases} 1 & \text{if } t \leq \frac{1}{2} \\ 0.16 \frac{\ln 2t + 1}{t - 0.15} \left(1 + \sqrt{1 + \frac{1.2(t - 0.15)}{\ln 2t + 1}} \right) & \text{if } t > \frac{1}{2}. \end{cases}$$

Note that this approximation up to the values of the constants coincides with the functions that define bounds in Chapter 4.

Table 13.1. Number of independent coordinates and effective VC dimension obtained from the measurements of learning machines

Number of independent coordinates:	40	30	20	10
Effective VC dimension:	40	31	20	11

13.2.4 Experiments on Measuring the Effective VC Dimension

The following experiments with measuring the effective VC dimension show that the obtained values provide a good estimate of the VC dimensions.

Example 1. Consider the following classifier: The classifier maps n -dimensional vectors x into n -dimensional vectors z using some degenerate linear transformation where only m coordinates of input vector x are linearly independent. In Z space the classifier constructs a linear decision rule.

For such a machine the VC dimension is equal to r_n , the number of linear independent coordinates. Using the described measuring method (with $\gamma = 0$ in functional (13.26)), we obtained results presented in Table 13.1.

Table 13.1 describes the experiments with four different machines that have the same dimensionality of the input space $n = 50$ and different number of linear independent coordinates in Z space (40, 30, 20, and 10).

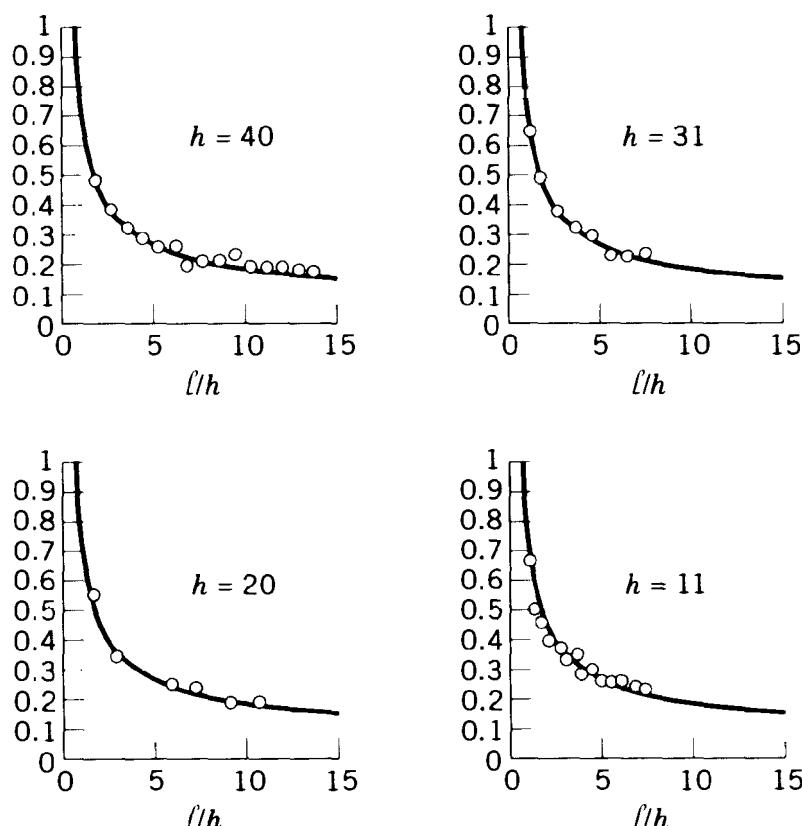
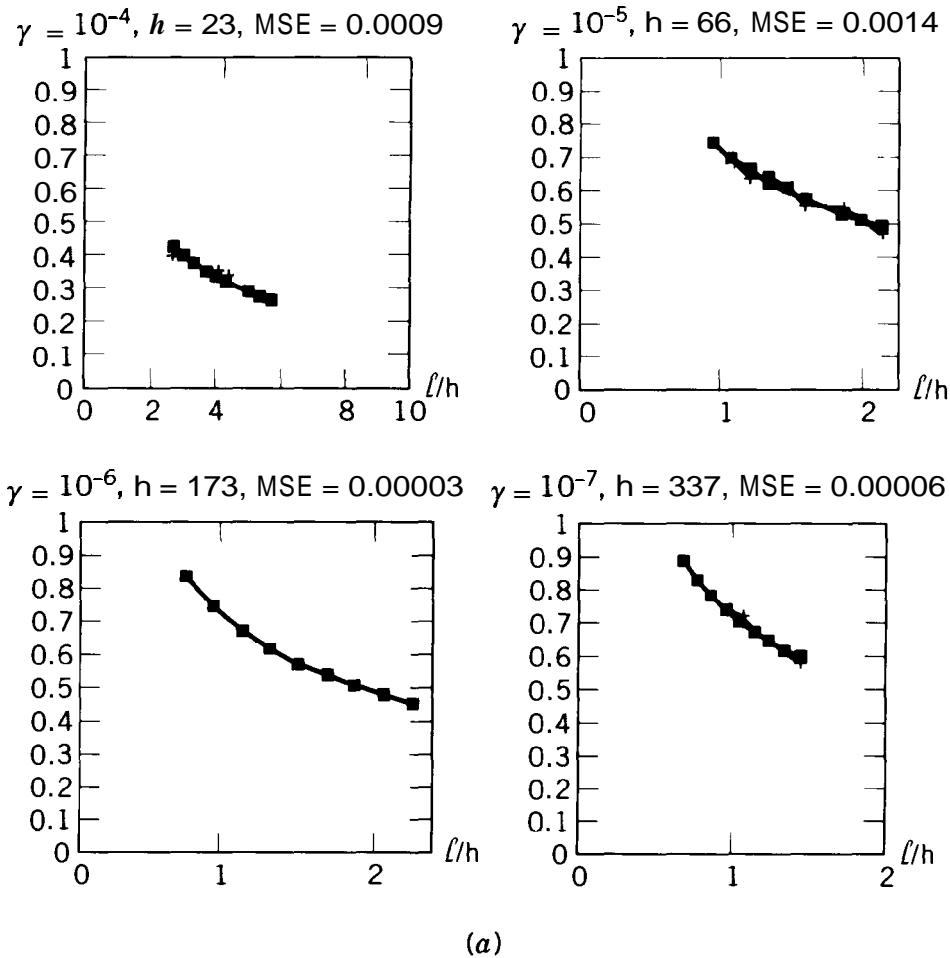


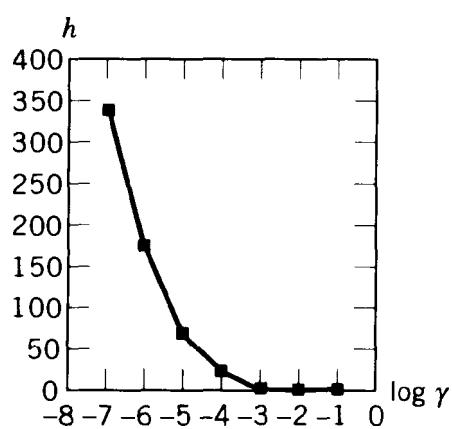
FIGURE 13.6. The best fit to the universal curve in experiments for estimating the VC dimension of machines with degenerating mapping.

Figure 13.6 shows the best fit to the universal curve for this experiment to the fitting curve $\Phi\left(\frac{\ell}{h}\right)$.

Example 2. Let us estimate the effective VC dimension of the SV machine



(a)



(b)

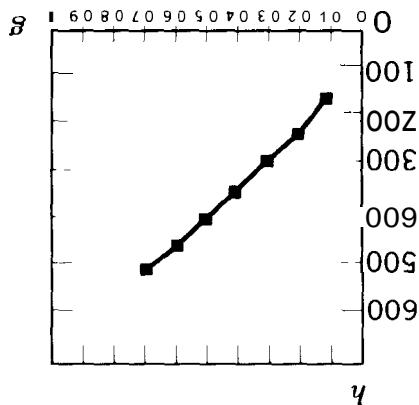
FIGURE 13.7. (a) The best fit to the universal curve in experiments for estimating the effective VC dimension of the SV machine with an RBF kernel for different regularization parameters γ . (b) The estimate of effective VC dimension of the SV machine with a fixed RBF kernel as the function of values γ .

database.

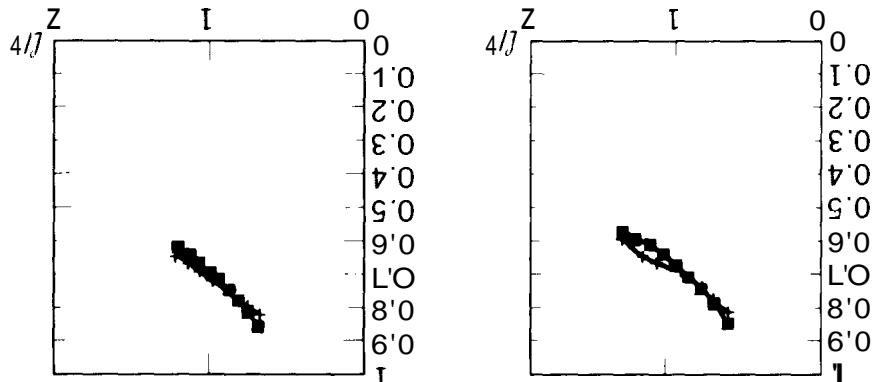
The experiment was conducted using the U.S. Postal Service databases γ . The effective VC dimension for different regularization parameters g and estimate the effective VC dimension for the parameter of width 256-dimensional space. In these experiments we fix the parameter of width with the radial basis function kernel $K(x, y) = \exp\{-g^2(x - y)^2\}$ drawn in

FIGURE 13.8. (a) The best fit in experiments for estimating the effective VC dimension of the SV machine with an RBF kernel for different parameters width g and fixed regularization parameter γ . (b) The estimate of effective VC dimension depending on parameter g .

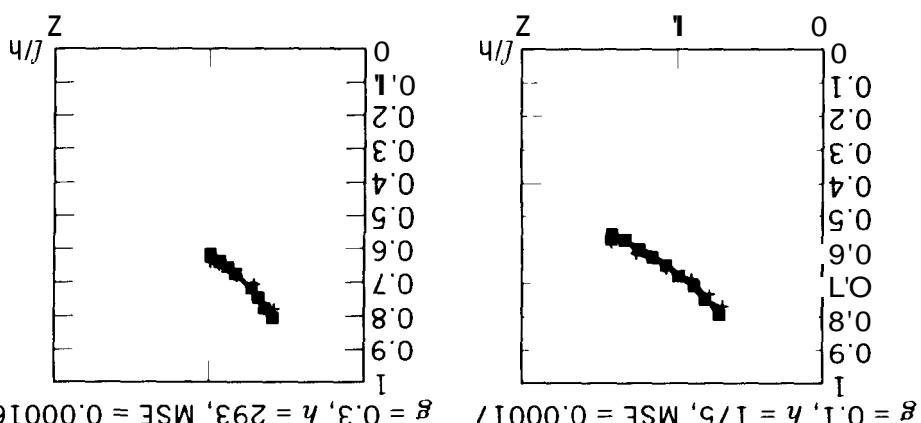
(b)



(a)



$$g = 0.5, h = \text{DOS}, \text{MSE} = 0.00033 \quad g = 0.6, h = 458.35, \text{SSE} = 0.00047$$



$$g = 0.1, h = 175, \text{MSE} = 0.00017 \quad g = 0.3, h = 293, \text{MSE} = 0.00016$$

Figure 13.7a shows the best fit to the universal curve in experiments for estimating the effective VC dimension of the SV machine with a fixed RBF kernel and different regularization parameters y . Figure 13.7b shows the function that defines the effective VC dimension depending on y . Every point on the plot is an average over 10 measurements. The following is shown: the value of parameter y , the estimated effective VC dimension h , and the mean-squared deviation of the measurements from the universal curve.

Figure 13.8a shows the best fit to the universal curve in experiments for estimating the effective VC dimension of the SV machine with RBF functions with different values of parameter g and fixed regularization parameter y . Figure 13.8b shows the estimate of effective VC dimension of the SV machine with an RBF kernel depending on parameter g . Every point on the plot is an average of more than 10 measurements. The following are shown: the values of parameter g , the estimated effective VC dimension h , and the mean-squared deviation of the measurements from the universal curve.

The experiments were conducted for the U.S. Postal Service database.

13.3 FUNCTION APPROXIMATION USING THE SV METHOD

Consider examples of solving the function approximation problem using the SV method. With the required level of accuracy ε we approximate one- and two-dimensional functions defined on a uniform lattice $x_i = ia/\ell$ by its values

$$(y_1, x_1), \dots, (y_\ell, x_\ell).$$

Our goal is to demonstrate that the number of support vectors that are used to construct the SV approximation depends on the required accuracy ε : The less accurate the approximation, the lower the number of support vectors needed.

In this section, to approximate real-valued functions we use linear splines with an infinite number of knots.

First we describe experiments for approximating the one-dimensional sinc function

$$f(x) = \frac{\sin(x - 10)}{x - 10} \quad (13.27)$$

defined on 100 points of the uniform lattice of the interval $0 \leq x \leq 20$.

Then we approximate the two-dimensional sinc function

$$f(x, y) = \frac{\sin \sqrt{(x - 10)^2 + (y - 10)^2}}{\sqrt{(x - 10)^2 + (y - 10)^2}} \quad (13.28)$$

defined on the points of the uniform lattice in $0 \leq x \leq 20$, $0 \leq y \leq 20$.

To construct the one-dimensional linear spline approximation we use the

kernel defined in Chapter 11, Section 11.7:

$$K(x, x_i) = 1 + x_i x + \frac{1}{2}|x - x_i|(x \wedge x_i)^2 + \frac{(x \wedge x_i)^3}{3}$$

We obtain an approximation in the form

$$y = \sum_{i=1}^N (\alpha_i^* - \alpha_i) K(x, x_i) + b,$$

where the coefficients α^* and α_i are the solution of the quadratic optimization problem defined in Chapter 11.

Figures 13.9 and 13.10 show approximations of the function (13.27) with different levels of accuracy. The circles on the figures indicate the support vectors. One can see that by decreasing the required accuracy of the approximation, the number of support vectors decreases.

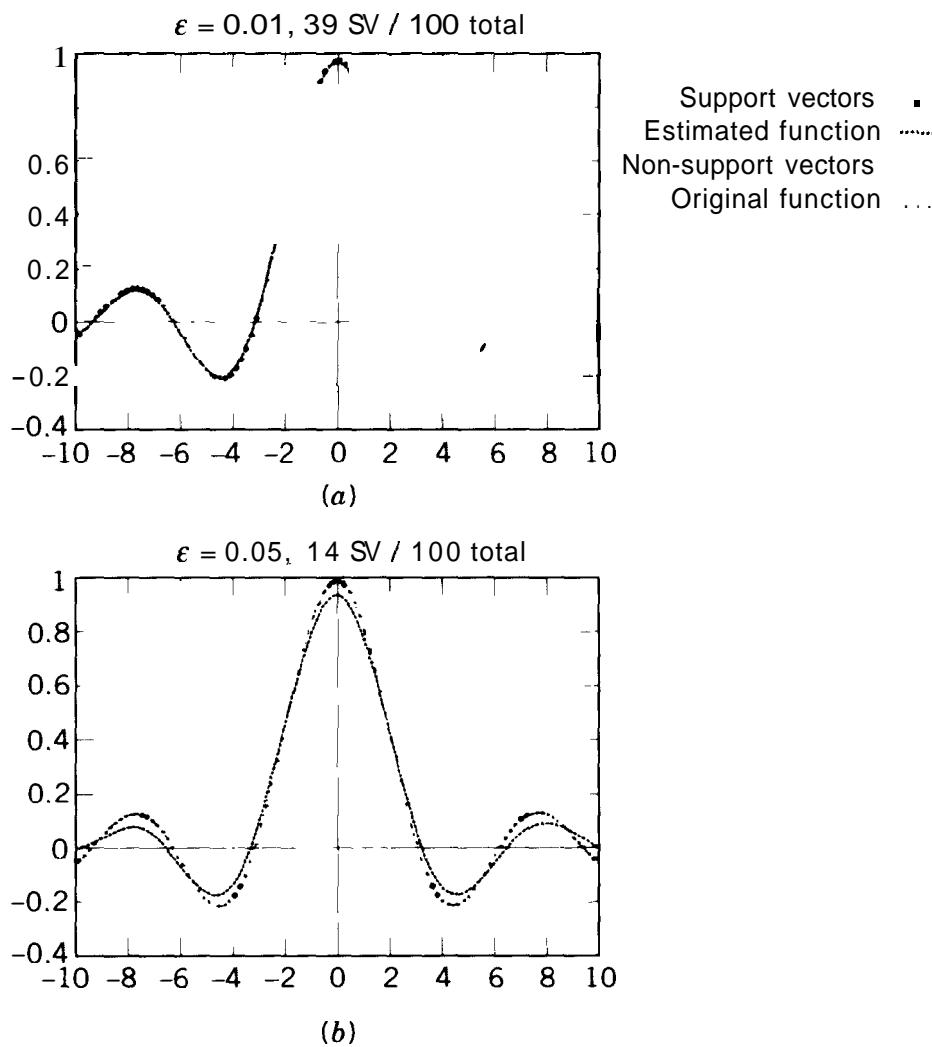


FIGURE 13.9. Approximations with a different level of accuracy requires a different number of support vectors: (a) 39 SV for $\epsilon = 0.01$, (b) 14 SV for $\epsilon = 0.05$.

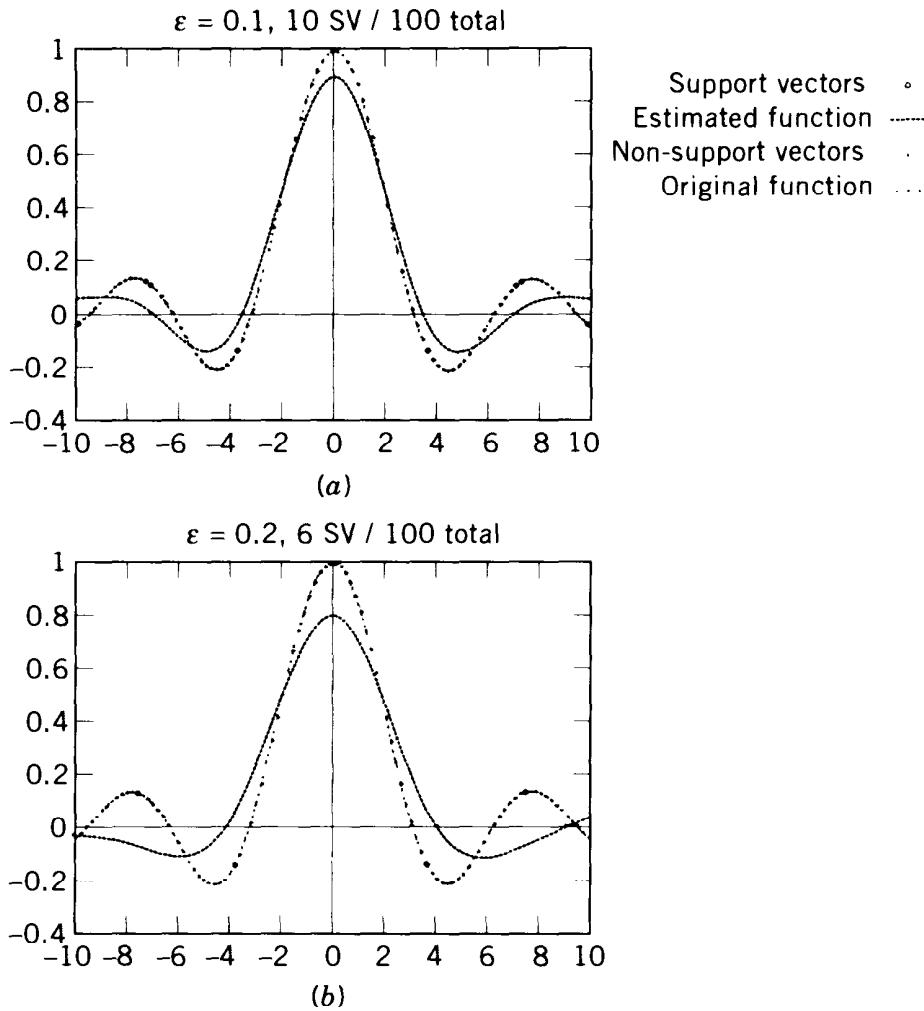


FIGURE 13.10. Approximations with a different level of accuracy requires a different number of support vectors: (a) 10 SV for $\epsilon = 0.1$, and (b) 6 SV for $\epsilon = 0.2$.

To approximate two-dimensional sinc function (13.28) we used the kernel

$$\begin{aligned} K(x, y; x_i, y_i) &= K(x, x_i)K(y, y_i) \\ &= \left(1 + xx_i + \frac{1}{2}|x - x_i|(\wedge x_i)^2 + \frac{(x \wedge x_i)^3}{3}\right) \\ &\quad \times \left(1 + yy_i + \frac{1}{2}|y - y_i|(y \wedge y_i)^2 + \frac{(y \wedge y_i)^3}{3}\right), \end{aligned}$$

which is defined by multiplication of the two one-dimensional kernels.

We obtain an approximation in the form

$$y = \sum_{i=1}^N (\alpha_i^* - \alpha_i) K(x, x_i) K(y, y_i) + b,$$

where coefficients a^* and a are defined by solving the same quadratic optimization problem as in the one-dimensional case.

Figure 13.11 shows the two-dimensional sinc function and its approximation with the required accuracy $\varepsilon = 0.03$ approximated using lattices with 400 points. Figure 13.12 shows the approximations obtained with the same accuracy $\varepsilon = 0.03$ using different number of grid points: 2025 in Fig. 13.12(a), and 7921 in Fig. 13.12(b). One can see that changing the number of grid points by a factor of 20 increases the number of support vectors by less than a factor of 2: 153 SV in Fig. 13.11(b), 234 SV in Fig. 13.12(a), and 285 SV in Fig. 13.12 (b).

13.3.1 Why Does the Value of ε Control the Number of Support Vectors?

The following model describes the mechanism of choosing the support vectors for function approximation using the SV machine with an ε -insensitive loss function.

Suppose one would like to approximate a function $f(x)$ with the accuracy ε —that is, to describe function $f(x)$ by another function $f^*(x)$ such that the function $f(x)$ is situated into the ε tube of $f^*(x)$. To construct such a function let us take an elastic ε tube (a tube that tends to be flat) and put function $f(x)$ into the ε tube. Since the elastic tube tends to become flat, it will touch some points of function $f(x)$. Let us fasten the tube at these points. Then the axis of the tube defines the ε approximation $f^*(x)$ of the function $f(x)$, and coordinates of the points where the ε -tube touches the function $f(x)$ define the support vectors. The kernel $K(x_i, x_j)$ describes the law of elasticity.

Indeed, since the function $f(x)$ is in the ε tube, there are no points of the function with distance of more than ε to the center line of the tube. Therefore the center line of the tube describes the required approximation.

To see that points which touch the ε -tube define the support vectors, it is sufficient to note that we obtained our approximation by solving an optimization problem defined in Chapter 11 for which the Kuhn–Tucker conditions hold. According to definition the support vectors are the ones for which in the Kuhn–Tucker condition the Lagrange multipliers are different from zero and hence the second multiplier must be zero. This multiplier defines the border points in an optimization problem of inequality type—that is, coordinates where the function $f(x)$ touches the ε tube. The wider the ε tube, the lower the number of touching points.

This model is valid for the function approximation problem in any number of variables.

Figure 13.13 shows the ε -tube approximation that corresponds to the case of approximation of the one-dimensional sinc function with accuracy $\varepsilon = 0.2$.

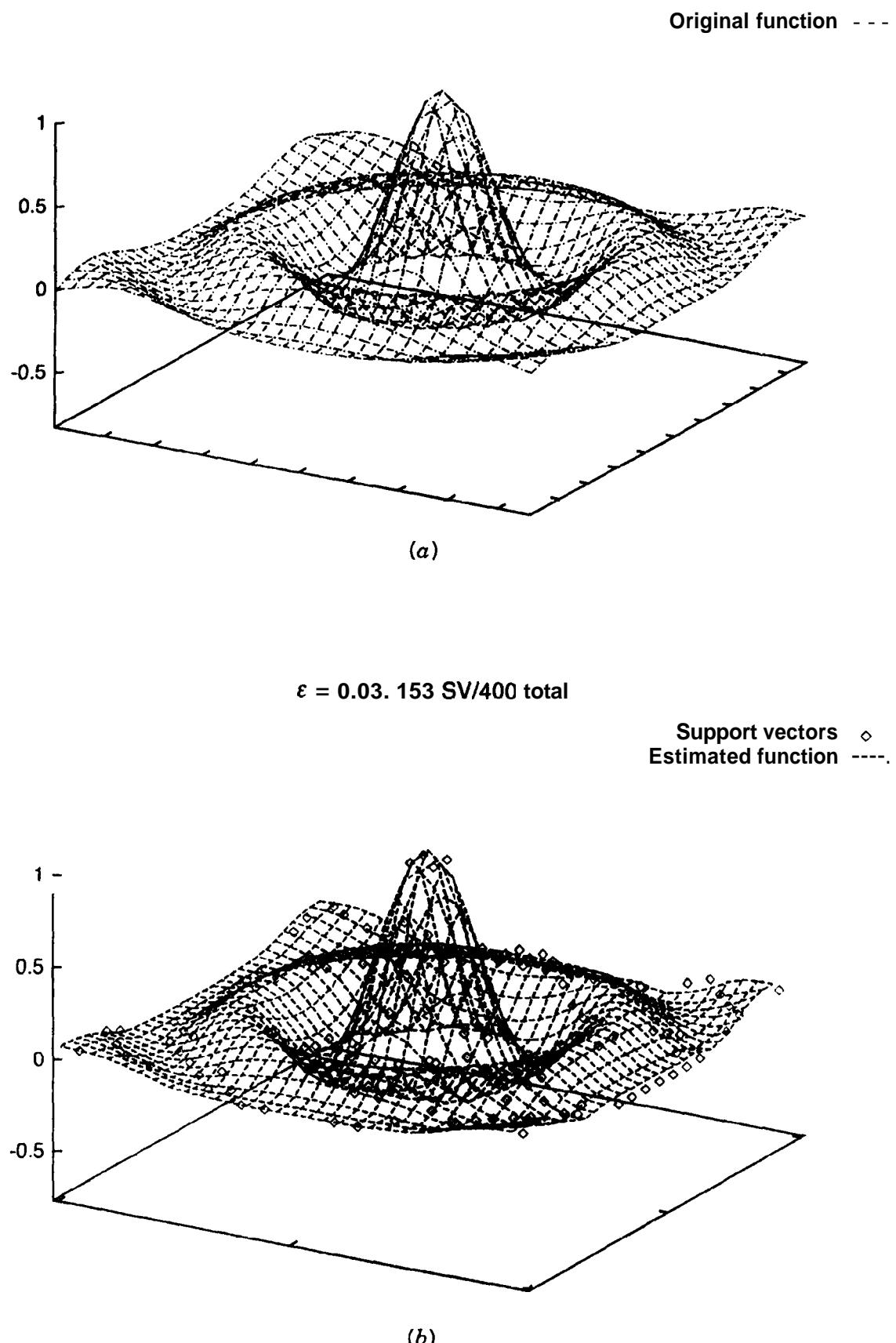
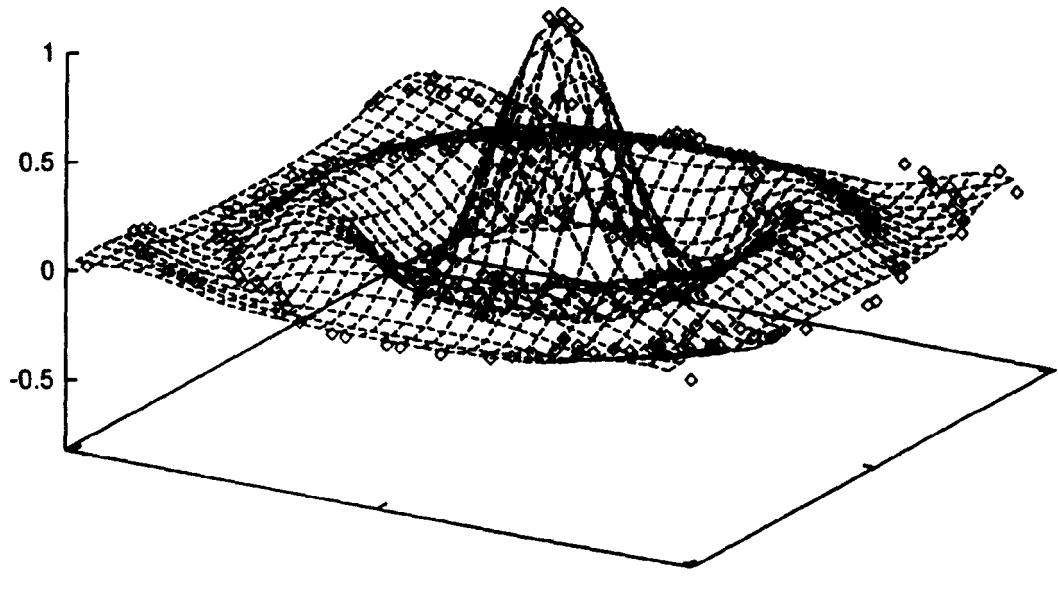


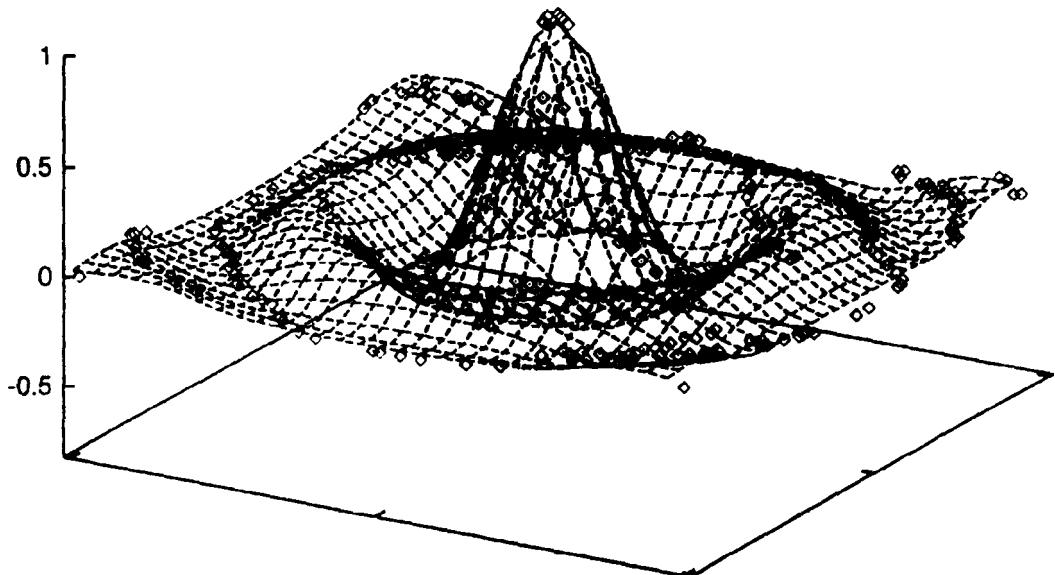
FIGURE 13.11. Two-dimensional sinc function (a) and its approximation with accuracy 0.03 obtained using 400 grid points. (b) The approximation was constructed on the basis of 153 SV (squares).

$\varepsilon = 0.03, 234 \text{ SV}/2025 \text{ total}$

Support vectors ◊
Estimated function ---



(a)

 $\varepsilon = 0.03, 285 \text{ SV}/7921 \text{ total}$


(b)

FIGURE 13.1.2. Approximations to two-dimensional sinc function defined on the lattices containing different numbers of grid points with the same accuracy $\varepsilon = 0.03$ does not require a big difference in the number of support vectors: (a) 234 SV for the approximation constructed using 2025 gridpoints, and (b) 285 SV for the approximation constructed using 7921 gridpoints.

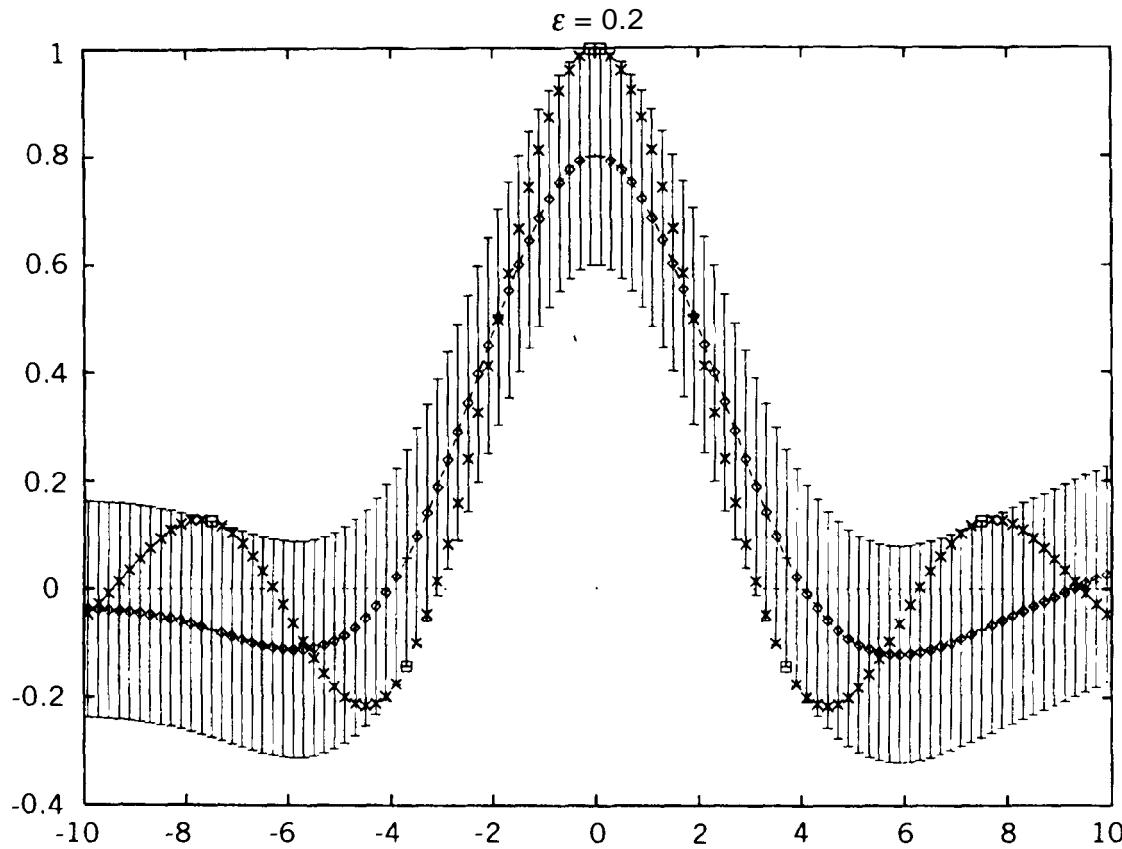


FIGURE 13.13. The ε -tube model of function approximation.

13.4 SV MACHINE FOR REGRESSION ESTIMATION

We start this section with simple examples of regression estimation tasks where regressions are defined by one- and two-dimensional sinc functions. Then we consider estimating multidimensional linear regression functions using the **SV** method. We construct a linear regression task that is favorable for a feature selection method and compare results obtained for a forward feature selection method with results obtained by the **SV** machine. Then we compare the **SV** regression method with new nonlinear techniques on three multidimensional artificial tasks suggested by J. Friedman and one multidimensional real-life (Boston housing) task (these tasks are usually used in benchmark studies of different regression estimation methods).

13.4.1 Problem of Data Smoothing

Let the set of data

$$(y_1, x_1), \dots, (y_\ell, x_\ell)$$

be defined by the one-dimensional sinc function on the interval $[-10, 10]$; the

values y_i are corrupted by noise with normal distribution

$$y_i = \frac{\sin x}{x} + \xi_i, \quad E\xi_i = 0, \quad E\xi_i^2 = \sigma^2.$$

The problem is to estimate the regression function

$$y = \frac{\sin x}{x}$$

from 100 observations on uniform lattice of interval $[-10, 10]$.

Figures 13.14 and 13.15 show the results of SV regression estimation experiments from data corrupted by different levels of noise. The rectangles in the figures indicate the support vectors. The approximations were obtained using linear splines with infinite number of knots.

Figures 13.16, 13.17, and 13.18 show approximation of the two-dimensional regression function

$$y = \frac{\sin \sqrt{x^2 + y^2}}{\sqrt{x^2 + y^2}}$$

defined on a uniform lattice on the interval $[-10, 10] \times [-10, 10]$. The approximations were obtained using a two-dimensional linear with an infinite number of knots.

13.4.2 Estimation of Linear Regression Functions

This section describes experiments with SV machines in estimating linear regression functions (Drucker et al., 1997).

We compare the SV machine to two different methods for estimating the linear regression function, namely the ordinary least square method (OLS) and the forward stepwise feature selection (FSFS) method.

Recall that the OLS method is a method that estimates the coefficients of a linear regression function by minimizing the functional

$$R(a) = \sum_{i=1}^{\ell} (y_i - (a * x_i))^2$$

The FSFS method is a method that first chooses one coordinate of the vector that gives the best approximation of data. Then it fixes this coordinate and adds a second coordinate such that these two define the best approximation of the data, and so on. One uses some technique (i.e., the functionals described in Section 13.1) to choose the appropriate number of coordinates.

We consider the problem of linear regression estimation from the data

$$(y_1, x_1), \dots, (y_\ell, x_\ell)$$

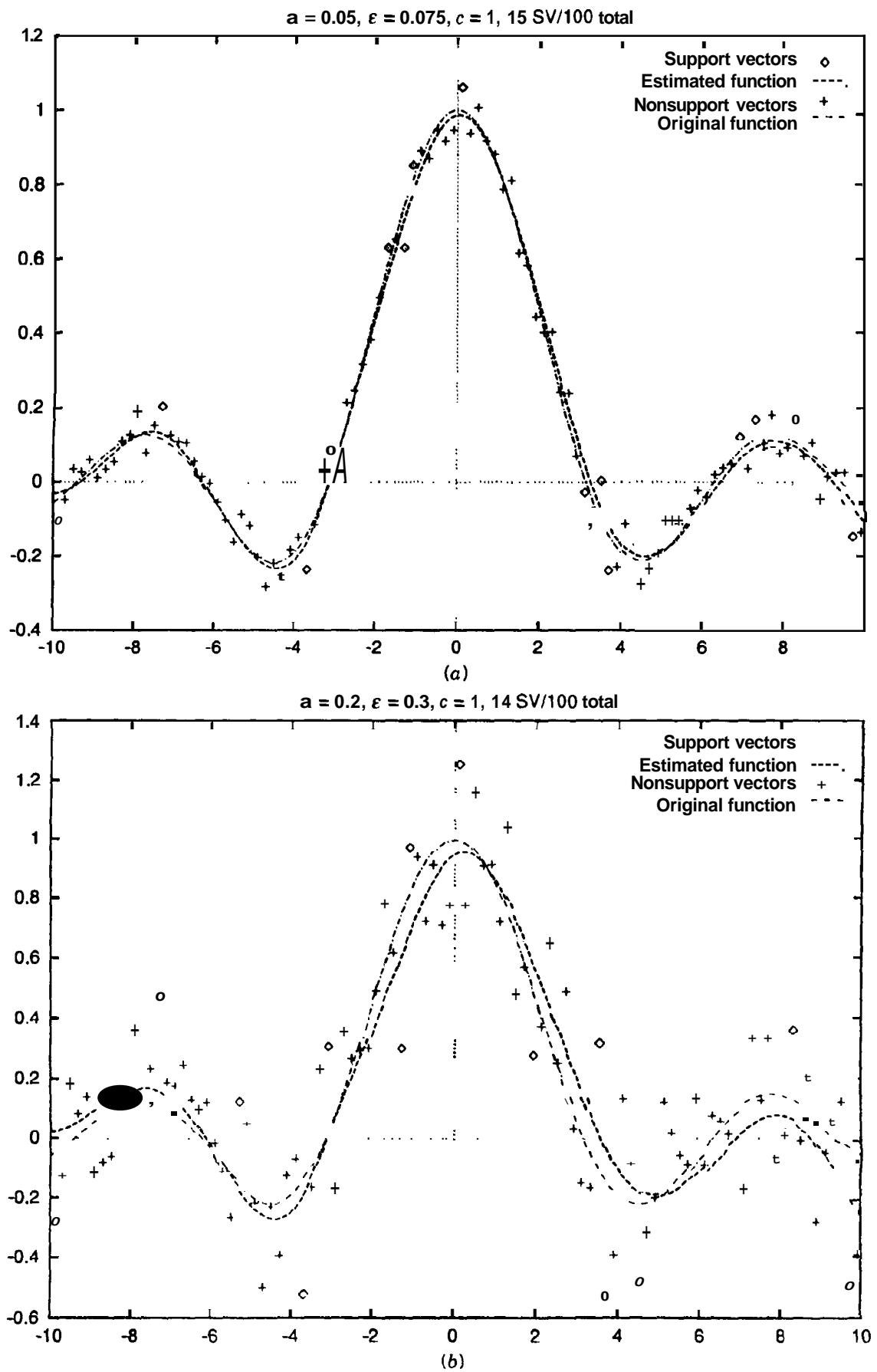


FIGURE 13.14. The regression function and its approximations obtained from the data with different levels of noise and different values ϵ : $\sigma = 0.05$ and $\epsilon = 0.075$ in (a); $\sigma = 0.2$ and $\epsilon = 0.3$ in (b). Note that the approximations were constructed using approximately the same number of support vectors (14 in part (a) and 15 in part (b)).

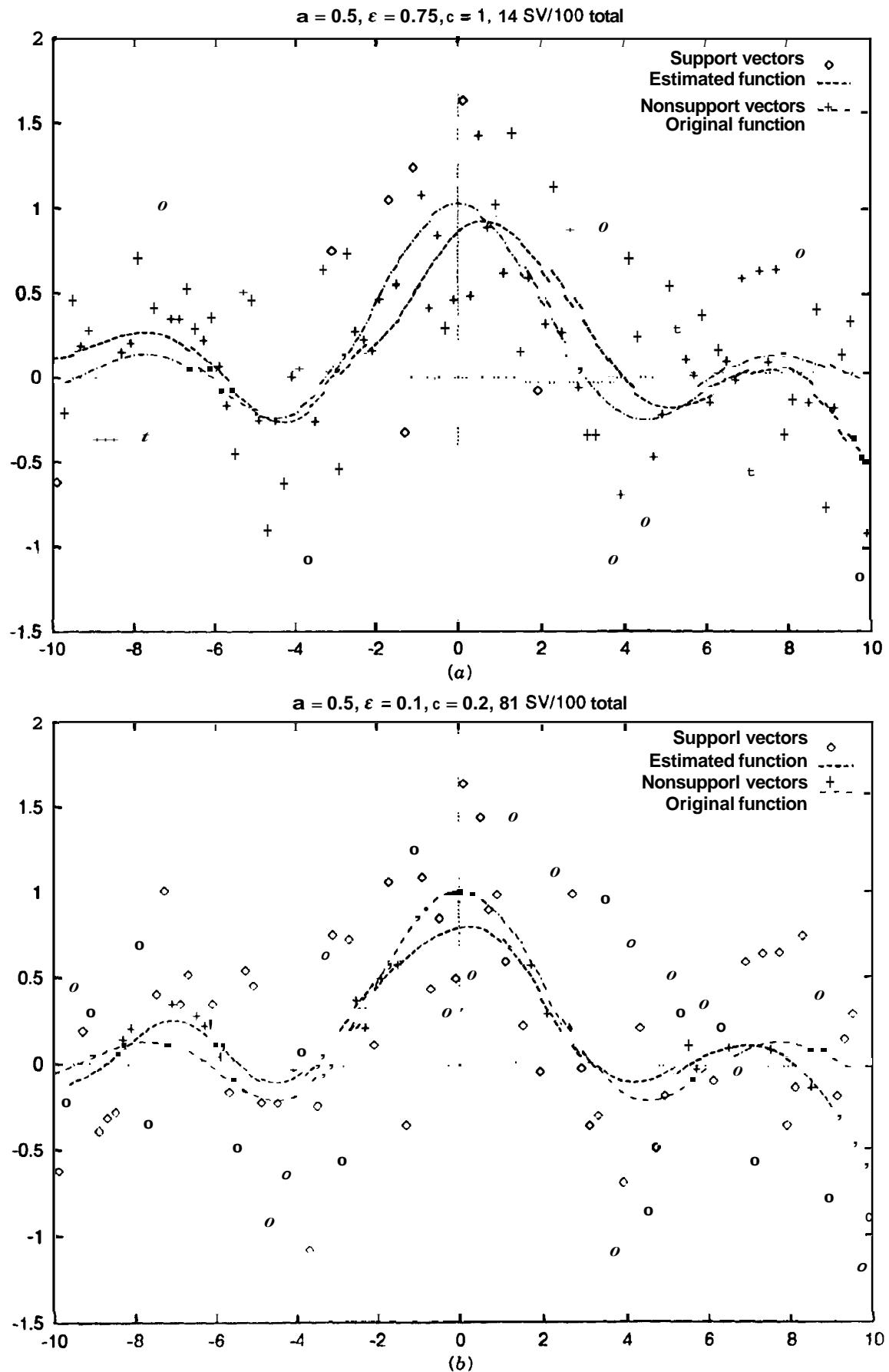
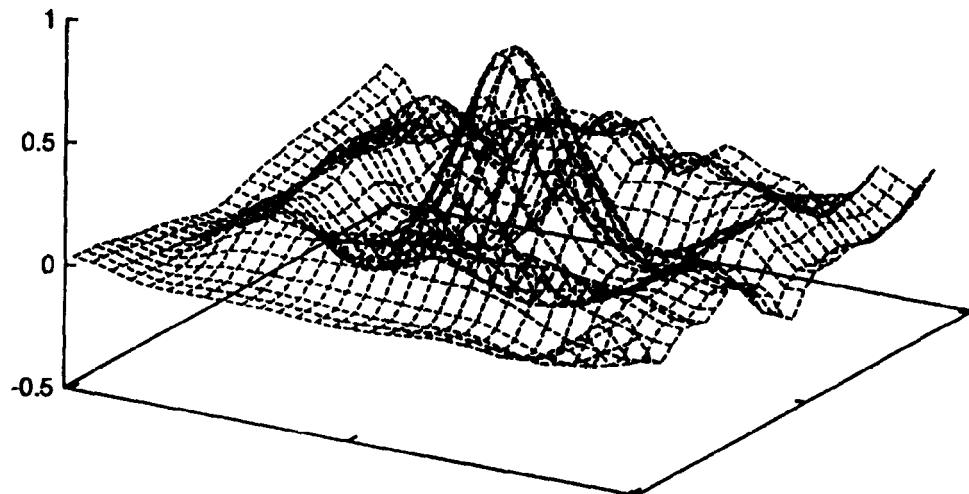


FIGURE 13.15. The regression function and its approximations obtained from the data with the same level of noise $\sigma = 0.5$ and different values of ϵ ($\epsilon = 0.25$ in (a) and $\epsilon = 0.15$ in (b)). Note that different value of ϵ imply a different number of support vectors in approximating function: 14 in (a) and 81 in (b)).

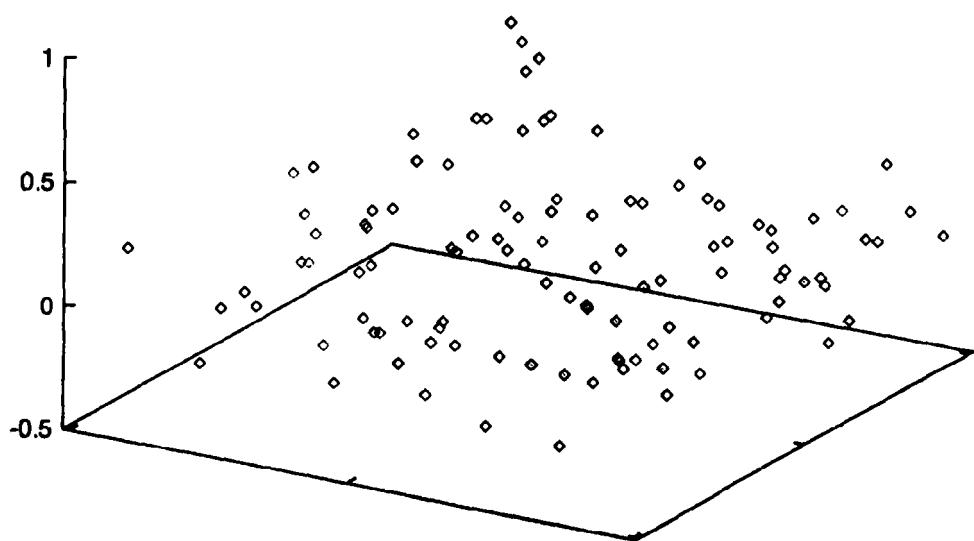
$\sigma = 0.1, \varepsilon = 0.15, 107 \text{ SV}/400 \text{ total}$

Estimated function ---



(a)

Support vectors ◊



(b)

FIGURE 13.16. (a) The approximation to the regression and (b) 107 support vectors, obtained from the data set of size 400 with noise $\alpha = 0.1$ and $\varepsilon = 0.15$

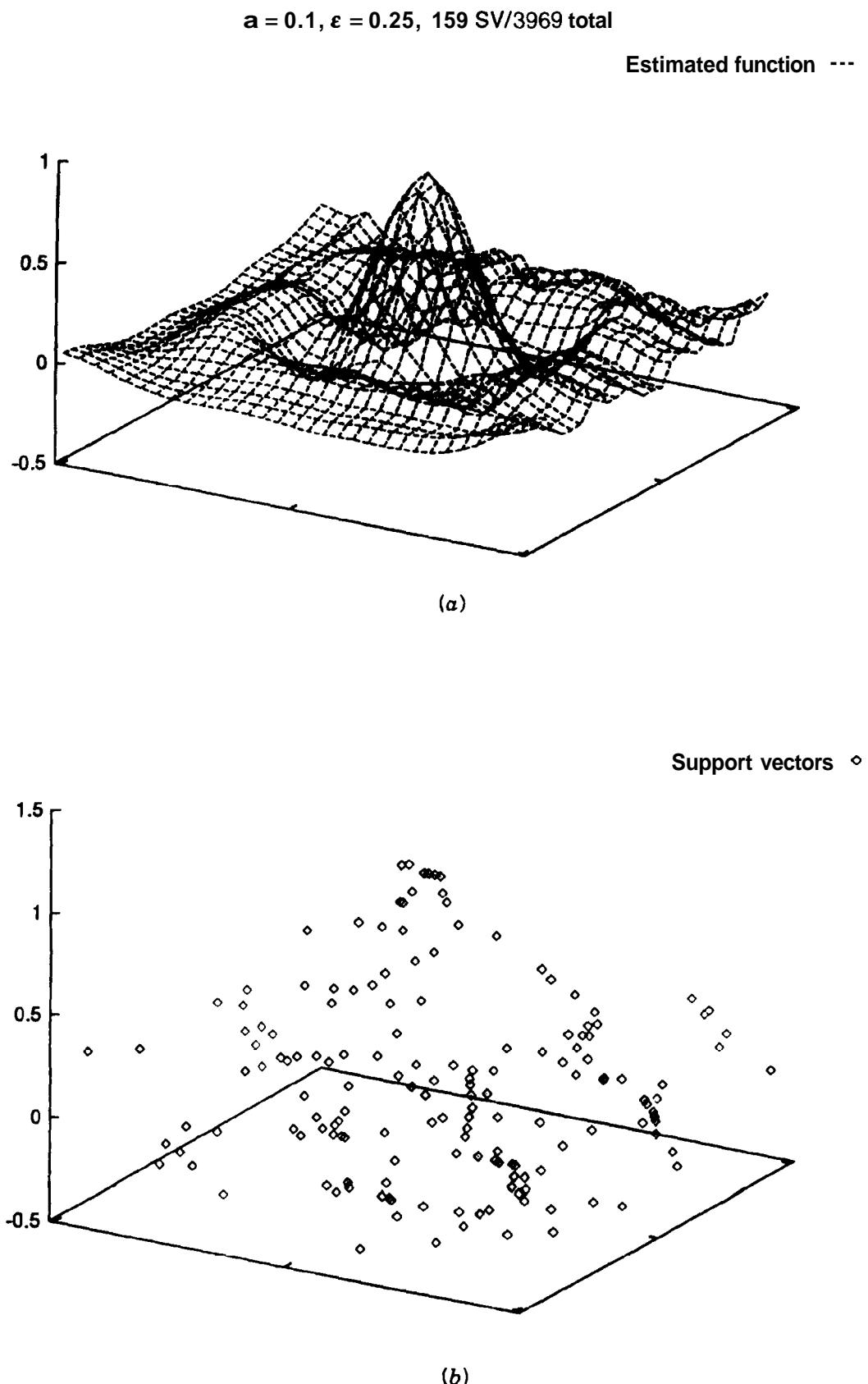


FIGURE 13.17. (a) The approximation to the regression and (b) 159 support vectors obtained from the data set of size 3969 with the same noise $\mathbf{a} = 0.1$ and $\epsilon = 0.25$.

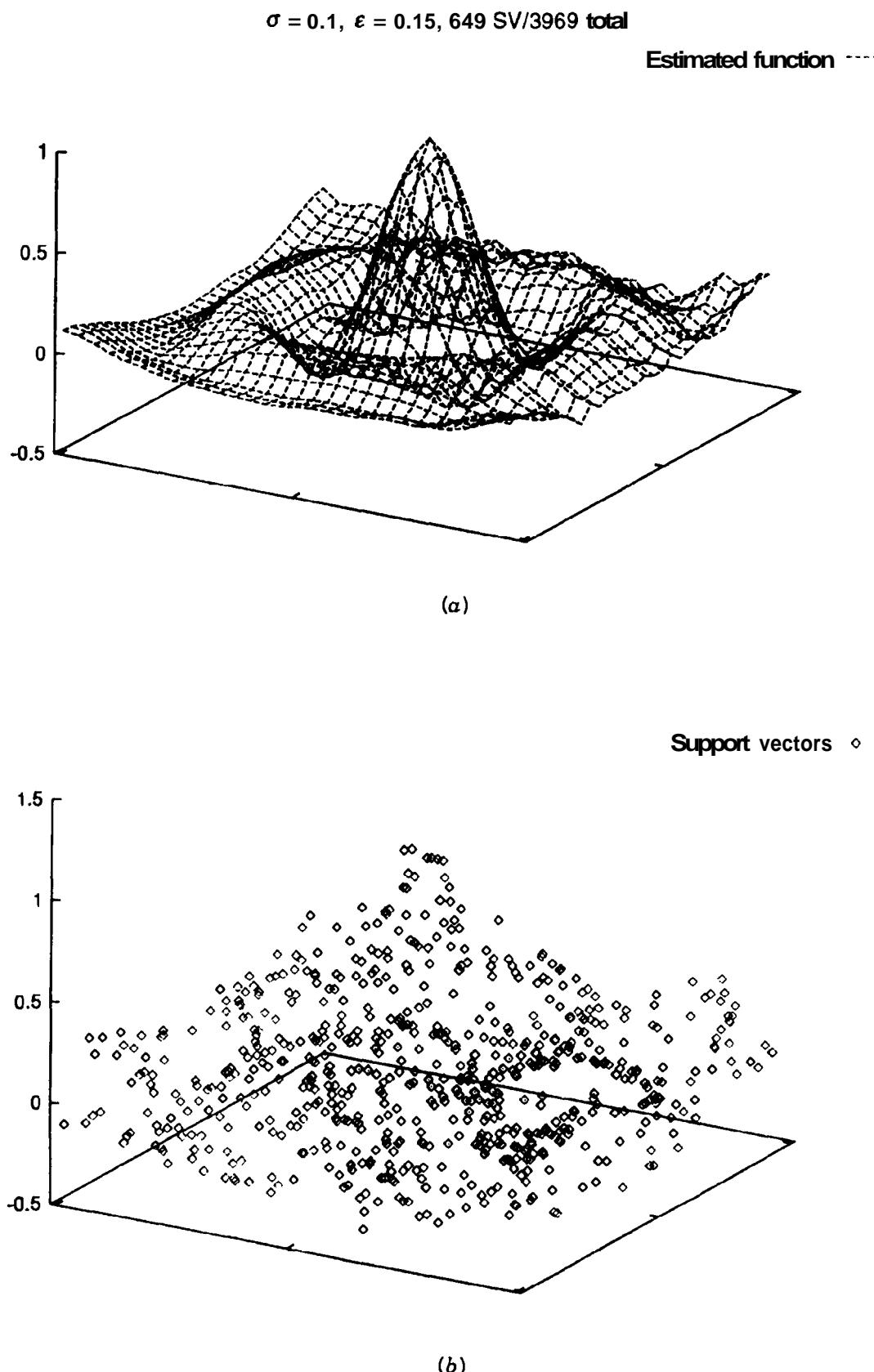


FIGURE 13.18. (a) the approximation to the regression and (b) 649 support vectors obtained from the data set of size 3969 with the same noise $\sigma = 0.1$ and $\epsilon = 0.15$.

Table 13.2. Result of comparison ordinary least-squared (OLS), forward step feature selection (FSFS), and support vector (SV) methods

SNR	Normal			Laplacian			Uniform		
	OLS	FSFS	SV	OLS	FSFS	SV	OLS	FSFS	SV
0.8	45.8	28.0	29.3	40.8	24.5	25.4	39.7	24.1	28.1
1.2	20.0	12.8	14.9	18.1	11.0	12.5	17.6	11.7	12.8
2.5	4.6	3.1	3.9	4.2	2.5	3.2	4.1	2.8	3.6
5.0	1.2	0.77	1.3	1.0	0.60	0.52	1.0	0.62	1.0

in 30-dimensional vector space $\mathbf{x} = (x^{(1)}, \dots, x^{(30)})$ where the regression function depends only on three coordinates

$$y(\mathbf{x}) = 2x_i^{(1)} + x^{(2)} + x_i^{(3)} + 0 \sum_{i=4}^{30} x^{(k)} \quad (13.29)$$

and the data are obtained as measurements of this function at randomly chosen points \mathbf{x} . The measurements are done with additive noise

$$y = y(x_i) + \xi$$

that is independent of x_i .

Table 13.2 shows that for large noise (small SNR) the SV regression gives results that are close to (favorable for this model) FSFS method, which is significantly better than the OLS method.

The experiments with the model

$$y_i = \sum_{i=1}^{30} x_i^{(k)} + \xi_i$$

demonstrated the advantage of SV technique for all levels of SNR defined in Table 13.2.

13.4.3 Estimation of Nonlinear Regression Functions

For these regression estimation experiments we chose regression functions suggested by Friedman (1991) that were used in many benchmark studies:

1. Friedman model #1 considered the following function of 10 variables

$$y = 10 \sin(\pi x^{(1)} x^{(2)}) + 20(x^{(3)} - 0.5)^2 + 10x^{(4)} + 5x^{(5)} + \xi. \quad (13.30)$$

This function, however, depends on only five variables. In this model the 10 variables are uniformly distributed in $[0, 1]$ and the noise is normal with parameters $N(0, 1)$.

2. Friedman model #2,

$$y = \sqrt{(x^{(1)})^2 + [x^{(2)}x^{(3)} - 1/(x^{(2)}x^{(3)})]^2},$$

has four independent variables uniformly distributed in the following region:

$$\begin{aligned} 0 &\leq x^{(1)} \leq 100, \\ 40\pi &\leq x^{(2)} \leq 560\pi, \\ 0 &\leq x^{(3)} \leq 1, \\ 1 &\leq x^{(4)} \leq 11. \end{aligned} \tag{13.31}$$

The noise is adjusted for a 3:1 SNR.

3. Friedman model #3 also has four independent variables

$$y = \tan^{-1} \left[\frac{x^{(2)}x^{(3)} - 1/x^{(2)}x^{(4)}}{x^{(1)}} \right] + \xi \tag{13.32}$$

that are uniformly distributed in the same region (13.31). The noise was adjusted for a 3:1 SNR.

In the following, we compare the SV regression machine with the advanced regression techniques called bagging (Breiman, 1996) and AdaBoost[†] (Freund and Schapire, 1995), which construct committee machines from the solutions given by regression trees. The experiments were conducted using the same format (Drucker, 1997; Drucker et al., 1997).

Table 13.3 shows results of experiments for estimating Friedman's functions using bagging, boosting, and a polynomial ($d=2$) SV methods. The experiments were conducted using 240 training examples. Table 13.3 shows

Table 13.3. Comparison of bagging and boosted regression trees with SV regression trees with SV regression in solving three Friedman tasks

	Bagging	Boosting	SV
Friedman #1	2.2	1.65	0.67
Friedman #2	11.463	11.684	5.402
Friedman #3	0.0312	0.0218	0.026

[†] AdaBoost algorithm was proposed for pattern recognition problem. It was adopted for regression estimation by H. Drucker (1997).

Table 13.4. Performance of Boston housing task for different methods

Bagging	Boosting	SV
12.4	10.7	7.2

Table 13.5. Performance of Boston housing task for different methods

MARS-I	MARS-3	POLYMARS
14.37	15.91	14.07

an average of more than 10 runs of the model error (mean square deviation between real regression function and obtained approximation).

Table 13.4 shows performance obtained from the Boston housing task where 506 examples of 13-dimensional real-life data were used as follows: 401 randomly chosen examples for the training set, 80 for the validation set and 25 for test set. Table 13.4 shows results of averaging more than 100 runs. The SV machine constructed polynomials (mostly of degree 4 and 5) chosen on the basis of validation set. For the Boston housing task, the performance shows the mean squared difference between predicted and actual values y on the test set.

Table 13.5 shows performance of the classical statistical methods for solving the Boston housing task: MARS1 (multivariate adaptive regression spline, linear—Friedman, 1991), MARS3 (multivariate adaptive regression spline, cubic), and POLYMARS (MARS-type method) reported by Stone et al. (1997). The direct comparisons, however, could not be done because the experiments were conducted under different formats: 303 random chosen data were used as the training set and 202 as the test set. The performance shows the mean squared difference between predicted and actual values y on the test set.

13.5 SV METHOD FOR SOLVING THE POSITRON EMISSION TOMOGRAPHY (PET) PROBLEM

In this section we consider the PET problem as an example of a solution of a linear operator equation using the SV technique.

13.5.1 Description of PET

Positron emission tomography is a medical diagnostic technique that involves the reconstruction of radio activity within the body following the injection or inhalation of a tracer labeled with a positron emitting nuclide. The mechanism of PET is the following: During the disintegration of the radioactive nucleus

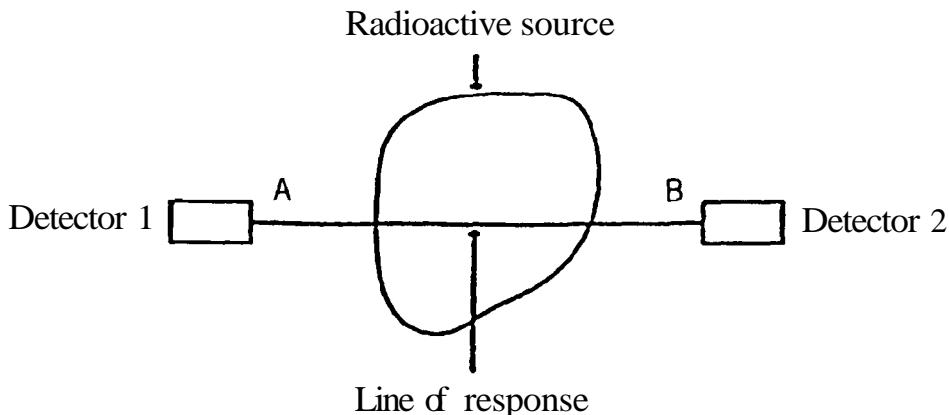


FIGURE 13.19. Scheme of PET measurements.

collected in the body, positrons are emitted. These positrons collide with nearby electrons, resulting in the annihilation of the electron and positron and the emission of two gamma rays in opposite directions. Figure 13.19 shows two directions on opposite sides of a radioactive source. From each point within the source, a gamma ray pair can be emitted in any direction. Two-dimensional PET, however, only takes into account the rays that belong to one fixed plane. In this plane if a gamma ray hits a detector 1 and then within a small time interval another gamma ray hits detector 2, then it is known that an emission must have accrued from a point somewhere along the line A – B joining these two detectors, the so-called line of response. This event is called a *coincidence*. The total number of coincidences for this pair of detectors is proportional to the integral of the tracer concentration along the line A – B. In order to obtain regional (in the plane) information about the tracer distribution, a large number of detector pairs with lines of response at many different angles are given. The set of all detector pairs whose lines of response are at a given angle μ form a μ projection. The set of all projections form a sinogram. Figure 13.20 illustrates two projections, each with six lines of response, and the corresponding sinogram.

Typically there are between 100 and 300 of these projection angles, μ_j , with each projection having between 100 and 200 lines of response m_i . This gives

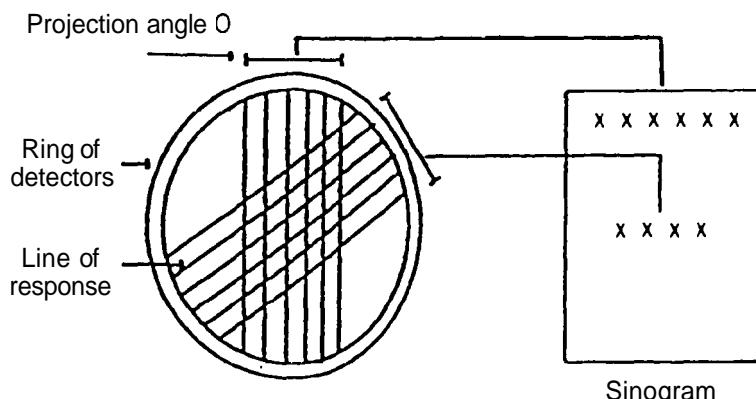


FIGURE 13.20. Scheme of data collection in the 2D PET problem.

between 10,000 and 60,000 of lines of response, each with a corresponding recorded number of coincidences $p(m_k, \mu_k)$. Therefore we are given ℓ triplets $m_k, \mu_k, p(m_k, \mu_k)$ called observations. The problem is given the set of observations (sinogram) to reconstruct the density of nuclear concentration within a given plane of the body.

13.5.2 Problem of Solving the Radon Equation

Consider Fig. 13.21, which shows a line of response inclined at the angle μ to the y axis and at a distance m from the origin. Let the circle that contains the detectors have radius 1. Suppose that coincidence count $p(m, \mu)$ is proportional to the integral of the concentration function $f(x, y)$ along the line defined by a pair m, μ . The operator, called the Radon transform operator, defines the integral of $f(x, y)$ along any line

$$\mathcal{R}f(x, y) = \int_{-\sqrt{1-m^2}}^{\sqrt{1-m^2}} f(m \cos \mu + u \sin \mu, m \sin \mu - u \cos \mu) du = p(m, \mu), \quad (13.33)$$

where coordinates x and y along the line are defined by the equations

$$\begin{aligned} x &= m \cos \mu + u \sin \mu, \\ y &= m \sin \mu - u \cos \mu \end{aligned} \quad (13.34)$$

and the position of the line is defined by the parameters

$$-1 < m < 1, \quad 0 \leq \mu \leq \pi.$$

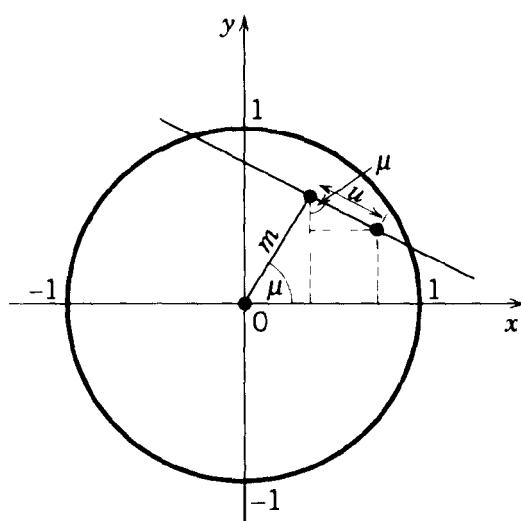


FIGURE 13.21. Parameters of the Radon equation.

The interval of integration is defined by

$$-a_m = -\sqrt{1-m^2} \leq u \leq \sqrt{1-m^2} = +a_m.$$

The main result of the theory of solving the Radon equation (given function $p(m, p)$, find the function $f(x, y)$ satisfying Eq. (13.33)) is that under wide conditions there exists the inverse operator

$$\mathcal{R}^{-1}\mathcal{R}[f(x, y)] = \mathcal{R}^{-1}p(m, \mu) = f(x, y).$$

In other words, there exists a solution to the Radon equation. However, finding this solution is an ill-posed problem.

Our goal is to find the solution to this ill-posed problem in a situation where function $p(m, \mu)$ is defined by its values p_k in a finite number ℓ of points m_k, μ_k , $k = 1, \dots, \ell$. Moreover, the data are not perfect: They are corrupted by some random noise

$$p_k = p(m_k, \mu_k) + \xi_k.$$

In other words, the problem is as follows: Given measurements

$$(p_1, m_1, \mu_1), \dots, (p_\ell, m_\ell, \mu_\ell),$$

find the solution to the Radon PET equation (1.33). Therefore we face the necessity of solving a stochastic ill-posed problem.

13.5.3 Generalization of the Residual Principle of Solving PET Problems

According to the theory for solving stochastic ill-posed problems described in the Appendix to Chapter 1 and in Chapter 7, in order to find the solution to the operator equation

$$\mathcal{A}f(t) = F(x) \quad (13.35)$$

using approximation $F_\ell(x)$ instead of the exact right-hand side $F(x)$ of equation (13.35), one has to minimize (in a given set of functions $\{f(t)\}$) the regularized functional

$$R = \|\mathcal{A}f(t) - F_\ell(x)\|^2 + \gamma W(f),$$

where $\gamma > 0$ is some regularization constant, and $W(f)$ is a regularizing functional.

In the PET problem, where we are given a finite number of measurements, usually one considers the following functional:

$$R(f) = \sum_{k=1}^{\ell} \left(p_k - \int_{-1}^1 f(m_k \cos \mu_k + u \sin \mu_k, m_k \sin \mu_k - u \cos \mu_k) du \right)^2 + \gamma W(f).$$

One also considers the set of piecewise constant or piecewise linear functions $\{f(x, y)\}$ in which one is looking for the solution.

In Section 11.11, when we considered the problem of solving an integral equation with an approximately defined right-hand side we discussed the idea of the residual method:

Suppose that solving the linear operator equation (13.35) with approximation $F_\ell(x)$ instead of $F(x)$, one has information about the accuracy of approximation

$$\Delta = \|F(x) - F_\ell(x)\|.$$

In this situation the residual method suggests that we choose the solution $f_\gamma(t)$ which minimizes the functional $W(f)$ and satisfies the constraint

$$\|\mathcal{A}f_\gamma(t) - F_\ell(x)\| \leq \Delta.$$

For the PET problem, one cannot evaluate the exact value of A ; the result of the measurement is a random event, the stochastic accuracy of which is characterized by the variance. The random variation in the number of coincidences along any line of response can be characterized as follows:

$$\varepsilon_k = \sqrt{p(m_k, \mu_k)}.$$

Therefore for the PET problem one can use a stronger regularization idea, namely to minimize the functional $W(f)$ subject to constraints

$$\left| p_k - \int_{-a_m}^{+a_m} f_\gamma(m_k \cos \mu_k + u \sin \mu_k, m_k \sin \mu_k - u \cos \mu_k) du \right| \leq \delta \varepsilon_k, \quad (13.36)$$

where $\delta > 0$ is some constant. In Chapter 11 we show that the SV method with an ε -insensitive loss function (with different ε_i for different vectors x_i) is the appropriate method for solving such problems. However, before applying the SV method for solving linear operator equations to the PET equation, let us briefly mention existing classical ideas for solving the PET problem.

13.5.4 The Classical Methods of Solving the PET Problem

The classical methods of solving the PET problem are based on finding the solution in a set of piecewise constant functions. For this purpose one can introduce the $n \times n = N$ pixel space where in any pixel one considers the value of the function to be constant. Let ℓ be the number of measurements. Then one can approximate the integral equation (13.35) by the algebraic equation

$$Ax = b, \quad x \geq 0, \quad (13.37)$$

where $A \in R^{\ell \times N}$ is a known matrix, $x \in R^N$ is a vector that defines the values of the approximating function in the set of pixels, and $b \in R^\ell$ is a vector that defines the number of coincidences along the lines of response.

The regularized method for the solution of Eq. (13.37) is the following:
Minimize the functional

$$R = (b - Ax)(b - Ax)^T + \gamma(x * x),$$

(here we use $W(x) = (x * x)$. One can use other regularizers as well.)

The residual principle for solving this problem led to the following statement: Choose a regularization parameter γ for which the solution x^* minimizes the functional (x, x) and satisfies the constraints

$$\left| b_k - \sum_{i=1}^{\ell} a_{i,k} x_i^* \right| \leq \epsilon_k,$$

where b_k is a coordinate of vector b , x_i^* is a coordinate of vector x^* , and $a_{i,k}$ is an element of matrix A .

The main problem with solving this equation is its size: As we mentioned, the size of M is $\approx 10,000\text{--}60,000$ observations and the number of parameters N to be estimated is $\approx 60,000$ ($N = 256 \times 256$).

This is a hard optimization problem. The classical methods of solving PET considered various ideas of solving this equation.

The main advantage of the SV method is that when using this method one does not need to reduce the solution of the PET problem to solving of the system of linear algebraic equations with huge number of variables.

13.5.5 The SV Method for Solving the PET Problem

We are looking for a solution of the PET problem in the set of two-dimensional spline functions with an infinite number of nodes. Below to simplify the formulas we consider (as in the classical case) piecewise constant functions (splines of order $d = 0$ with an infinite number of knots). One can solve this problem in other sets of functions, say by expansion on B splines or using an expansion on Gaussians (there is a good approximation to B splines; see Chapter 11, Section 11.7).

Thus, let us approximate the desired function by the expression

$$f(x, y) = \int_{-1}^1 \int_{-1}^1 \theta(x - t) \theta(y - \tau) \psi(t, \tau) dt d\tau, \quad (13.38)$$

where $\psi(t, \tau)$ is an appropriate function of Hilbert space.

Putting this function in the Radon equation, we obtain the corresponding regression problem in image space:

$$\begin{aligned} \mathcal{R} \int_{-1}^1 \int_{-1}^1 \theta(x - t) \theta(y - \tau) \psi(t, \tau) dt d\tau \\ = \int_{-1}^1 \int_{-1}^1 \mathcal{R} [\theta(x - t) \theta(y - \tau)] \psi(t, \tau) dt d\tau = p(m, \mu). \end{aligned} \quad (13.39)$$

Consider the following functions:

$$\begin{aligned}
 & \Phi(m_k, \mu_k, t, \tau) \\
 &= \mathcal{R}_{m_k, \mu_k} \theta(x - t) \theta(y - \tau) \\
 &- \int_{-a_{m_k}}^{+a_{m_k}} \theta(m_k \cos \mu_k + u \sin \mu_k - t) \theta(m_k \sin \mu_k - u \cos \mu_k - \tau) du.
 \end{aligned} \tag{13.40}$$

Using this function, we rewrite equality (13.39) as follows:

$$\int_{-1}^1 \int_{-1}^1 \Phi(m_k, \mu_k, t, \tau) \psi(t, \tau) dt d\tau = p(m_k, \mu_k), \quad k = 1, \dots, \ell.$$

Thus, we reduce the problem of solving the **PET** equation in a set of piecewise constant functions to the problem of regression approximation in the image space using the data p_i, m_i, μ_i , $i = 1, \dots, \ell$. We would like to find the function $\psi^*(t, \tau)$ that satisfies the conditions

$$\left| p_k - \int_{-1}^1 \int_{-1}^1 \Phi(m_k, \mu_k; t, \tau) \psi(t, \tau) dt d\tau \right| \leq \delta \varepsilon_k$$

and that has a minimal norm. We will use this function in Eq. (13.38) to obtain the solution of the desired problem.

To solve the **PET** problem using the SV technique, we construct two functions: the kernel function in the image space

$$K(m_i, \mu_i; m_j, \mu_j) = \int_{-1}^1 \int_{-1}^1 \Phi(m_i, \mu_i; t, \tau) \Phi(m_j, \mu_j; t, \tau) dt d\tau \tag{13.41}$$

and the cross-kernel function

$$\mathcal{K}(m_i, \mu_i, x, y) = \int_{-1}^1 \int_{-1}^1 \Phi(m_i, \mu_i, t, \tau) \theta(x - t) \theta(y - \tau) dt d\tau, \tag{13.42}$$

where the function $\Phi(m, \mu; t, \tau)$ is defined by expression (13.40).

To obtain these kernels in explicit form, we change the order of integration in (13.41) and (13.42) and take into account the following equality:

$$\begin{aligned}
 & \int_{-1}^1 \int_{-1}^1 \theta(x_i - t) \theta(x_j - t) \theta(y_i - \tau) \theta(y_j - \tau) dt d\tau \\
 &= (2 + \min(x_i, x_j))(2 + \min(y_i, y_j)).
 \end{aligned}$$

We obtain the following expression for the kernel function:

$$\begin{aligned}
 K(m_i, \mu_i; m_j, \mu_j) &= \int_{-1}^1 \int_{-1}^1 \Phi(m_i, \mu_i; t, \tau) \Phi(m_j, \mu_j; t, \tau) dt d\tau \\
 &= \int_{-1}^1 \int_{-1}^1 [\mathcal{R}_{m_i, \mu_i} \theta(x_i - t) \theta(y_i - \tau)] \times [\mathcal{R}_{m_j, \mu_j} \theta(x_j - t) \theta(y_j - \tau)] dt d\tau \\
 &= \mathcal{R}_{m_i, \mu_i} \mathcal{R}_{m_j, \mu_j} \left\{ \int_{-1}^1 \int_{-1}^1 \theta(x_i - t) \theta(x_j - t) \theta(y_i - \tau) \theta(y_j - \tau) dt d\tau \right\} \\
 &= \int_{-a_{m_i}}^{a_{m_i}} \int_{-a_{m_j}}^{a_{m_j}} [2 + \min \{(m_i \cos \mu_i + u_1 \sin \mu_i), (m_j \cos \mu_j + u_2 \sin \mu_j)\}] \\
 &\quad \times [2 + \min \{(m_i \sin \mu_i - u_1 \cos \mu_i), (m_j \sin \mu_j - u_2 \cos \mu_j)\}] du_1 du_2. \tag{13.43}
 \end{aligned}$$

We also obtain the following expression for the cross-kernel function:

$$\begin{aligned}
 \mathcal{K}(m_i, \mu_i, x, y) &= \int_{-1}^1 \int_{-1}^1 [\mathcal{R}_{m_i, \mu_i} \theta(x_i - t) \theta(y_i - \tau)] \times [\theta(x - t) \theta(y - \tau)] dt d\tau \\
 &= \mathcal{R}_{m_i, \mu_i} \left\{ \int_{-1}^1 \int_{-1}^1 \theta(x_i - t) \theta(y_i - \tau) \theta(x - t) \theta(y - \tau) dt d\tau \right\} \\
 &= \int_{-a_{m_i}}^{a_{m_i}} [2 + \min \{(m_i \cos \mu_i + u_1 \sin \mu_i), x\}] \\
 &\quad \times [2 + \min \{(m_i \sin \mu_i - u_1 \cos \mu_i), y\}] du_1. \tag{13.44}
 \end{aligned}$$

After elementary but cumbersome calculations, one can compute these piecewise polynomial integrals analytically.

Now to solve the PET problem on the basis of the SV technique, we need to do the following:

First, using kernel function (13.43) we need to obtain the SV solution for the regression approximation problem in image space. That is, we need to obtain the support vectors (m_k, μ_k) , $k = 1, \dots, N$, and the corresponding coefficients $\alpha_k^* - \alpha_k$, $k = 1, \dots, N$.

Second, using the cross-kernel function (13.44), we need to use the obtained support vectors and the obtained coefficients to define the desired approximation

$$f(x, y) = \sum_{k=1}^N (\alpha_k^* - \alpha_k) \mathcal{K}(m_k, \mu_k; x, y).$$

Note that for any support vector (m_k, μ_k) in image space there is a corre-

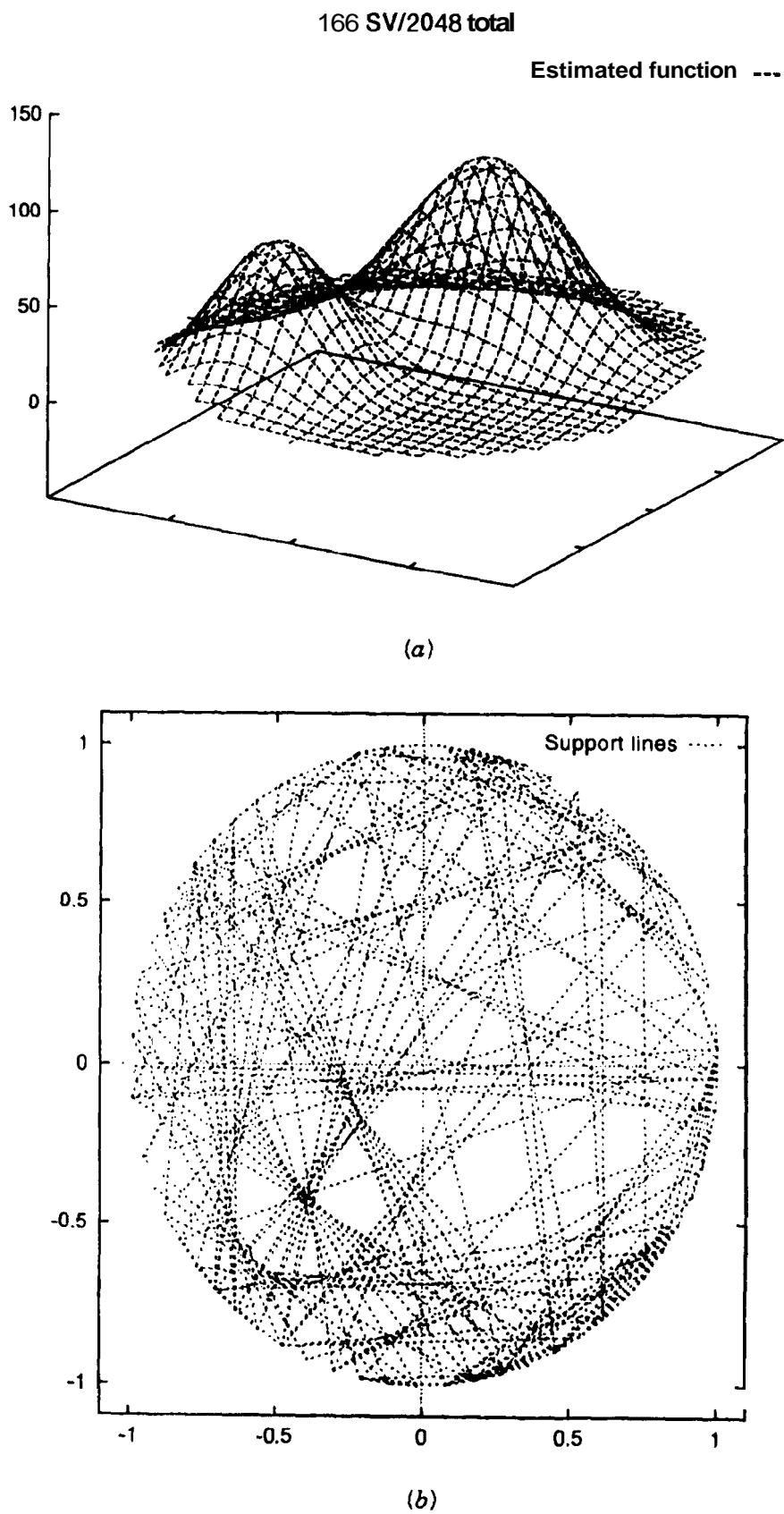


FIGURE 13.22. (a) Reconstructed image obtained on (b) the basis of 166 support lines.

sponding line of response in preimage space defined by the expression

$$\begin{aligned}x &= m_k \cos \mu_k + u \sin \mu_k, \\y &= m_k \sin \mu_k - u \cos \mu_k, \\-\sqrt{1 - m_k^2} &\leq u \leq \sqrt{1 - m_k^2}.\end{aligned}$$

Therefore in preimage space the expansion of the function on support vectors (m_k, μ_k) , $k = 1, \dots, N$, actually means the expansion of the desired solution on the lines of response.

Figure 13.22a shows a reconstructed image from 2048 observations in modeling the PET scan. The obtained spline approximation (of order $d = 0$) was constructed on the basis of 166 support vectors. Figure 13.22b shows the lines of response corresponding to the support vectors.

13.6 REMARK ABOUT THE SV METHOD

In the last two chapters we described some examples of applying the **SV** method to various function estimation tasks. We considered a wide range of tasks, starting with a relatively simple one (estimating indicator functions) and concluding with a relatively complex one (solving ill-posed operator equations based on measurements of its right-hand side).

In all dependency estimation tasks we tried, the very straightforward implementation of the **SV** approach demonstrated good results.

In the simplest dependency estimation problem—the pattern recognition problem—we obtained results that were not worse than the results obtained by the special state-of-the-art learning machines constructed for this specific problem.

In the function approximation tasks we were able both to construct approximations using a very rich set of functions and to control the trade-off between accuracy of approximation and complexity of the approximating function.

In examples of regression estimation the achieved advantage in accuracy compared to classical state-of-the-art methods sometimes was significant.

In the PET problem we did not create an intermediate problem—the pixels representation. We solved this problem in functional space.

In solving all the described examples we did not use any engineering. It is known, however, that special tailoring of the methods for the problem at hand is an important source of performance improvement. The **SV** approach has a rich opportunity for such tailoring.



STATISTICAL FOUNDATION OF LEARNING THEORY

Part III studies uniform laws of large numbers that make generalization possible.

NECESSARY AND SUFFICIENT CONDITIONS FOR UNIFORM CONVERGENCE OF FREQUENCIES TO THEIR PROBABILITIES

The last three chapters of this book studies the convergence of empirical processes. In this chapter we derive the necessary and sufficient conditions of uniform two-sided convergence of the frequencies to their probabilities.

According to the classical Bernoulli's theorem the frequency of any random event A converges to the probability of this event in the sequence of independent trials.

In the first part of this book it was shown that it is important to have convergence in probability simultaneously for all events $A \in S$ of a given set of events S, the so-called case of *uniform convergence of the frequencies to their probabilities over the given set of events*. In other words, it is important to guarantee the existence of uniform convergence in probability of the averages to their mathematical expectations over the given set of indicator functions, that is,

$$\sup_{\alpha \in \Lambda} \left| \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i \alpha) - \int Q(z, \alpha) dP(z) \right| \xrightarrow[\ell \rightarrow \infty]{} 0.$$

However, to show the relation of the results obtained here to the problem of the theoretical statistics (discussed in Chapter 2) we shall use the classical terminology, which is a little different from the terminology used in the first part of the book. Instead of the set of indicator functions $Q(z, a)$, $a \in A$, we shall consider the set S of events $A(\alpha) = \{z : Q(z, a) > 0\}$, $a \in A$, and instead of conditions for uniform convergence of the averages to their mathematical expectation over a given set of indicator functions we shall consider condi-

tions for uniform convergence of the frequencies $\nu(A)$ to their probabilities $P(A)$ over a given set S of events A :

$$\sup_{A \in S} |\nu(A) - P(A)| \xrightarrow[\ell \rightarrow \infty]{} 0.$$

It is clear that these two problems of uniform convergence are completely equivalent.

To stress that this part of the book has more general goals than foundations of the learning theory (the results obtained here actually belongs to foundations of the theoretical statistics), we also change the notations of the space. Instead of the space Z , which has a specific structure in the learning theory, we consider an abstract space X .

14.1 UNIFORM CONVERGENCE OF FREQUENCIES TO THEIR PROBABILITIES

Let X be a set of elementary events and let $P(x)$ be a probability measure defined on this set. Let S be some collection of random events—that is, a collection of subsets A of the set X measurable with respect to probability measure $P(x)$ (S is included in a σ algebra of random events but does not necessarily coincide with it).

Denote by $X(\ell)$ the space of samples from X of size ℓ . Because this sample was obtained in iteration of independent trials with the same distribution, we formalize by the assignment of the product measure on $X(\ell)$.

For any sample

$$X^\ell = x_1, \dots, x_\ell$$

and any event $A \in S$ the frequency of appearance of the event A is determined. It is equal to the number $n(A)$ of elements of the sample which belong to the event A , divided by the size ℓ of the sample:

$$\nu(A; X^\ell) = \nu(A; x_1, \dots, x_\ell) = \frac{n(A)}{\ell}.$$

Bernoulli's theorem asserts that for fixed event A the deviation of the frequency from probability converges to zero (in probability) when the sample size increases; that is, for any A and any ε the convergence

$$P\{|P(A) - \nu(A; X^\ell)| > \varepsilon\} \xrightarrow[\ell \rightarrow \infty]{} 0$$

holds true.

In this chapter we are interested in the maximal (for the given set S) deviation of the frequency from its probability:

$$\pi^S(X^\ell) = \sup_{A \in S} |P(A) - \nu(A; X^\ell)|.$$

The value $\pi^S(X^\ell)$ is a function of the point X^ℓ in the space $X(\ell)$. We will assume that this function is measurable with respect to the measure on $X(\ell)$ —that is, that $\pi^S(X^\ell)$ is a random variable.

We say that the frequencies of events $A \in S$ converge (in probability) to their probabilities uniformly on the set S if the random variable $\pi^S(X^\ell)$ tends in probability to zero as the value ℓ increases.

This chapter is devoted to the estimation of the probability of the event

$$\left\{ \pi^S(X^\ell) > \varepsilon \right\}$$

and determining the conditions when for any $\varepsilon > 0$ the equality

$$\lim_{\ell \rightarrow \infty} P \left\{ \pi^S(X^\ell) > \varepsilon \right\} = 0$$

holds true.

14.2 BASIC LEMMA

We start by considering an important lemma. We are given a sample of size 2ℓ :

$$X^{2\ell} = x_1, \dots, x_\ell, x_{\ell+1}, \dots, x_{2\ell}.$$

For the event $A \in S$ we calculate from the first half-sample

$$X_1^\ell = x_1, \dots, x_\ell$$

the frequency

$$\nu_1(A; X^{2\ell}) = \frac{n(A; X_1^\ell)}{\ell},$$

and from the second half-sample

$$X_2^\ell = x_{\ell+1}, \dots, x_{2\ell}$$

we calculate the frequency

$$\nu_2(A; X^{2\ell}) = \frac{n(A; X_2^\ell)}{\ell}$$

Consider the deviations between these two frequencies:

$$\rho(A; X^{2\ell}) = |\nu_1(A; X^\ell) - \nu_2(A; X^\ell)|.$$

Denote the maximal deviation over the given set S of events A by

$$\rho^S(X^{2\ell}) = \sup_{A \in S} \rho(A; X^{2\ell}).$$

We suppose that the function $\rho^S(X^{2\ell})$ is measurable.

Thus for the given set S of events we have to construct two random variables: the random variable $\pi^S(X^\ell)$ and the random variable $\rho^S(X^{2\ell})$.

As will be shown in subsequent sections, it is possible both to upperbound the distribution function of the random variable $\rho^S(X^{2\ell})$ and low bound this distribution function. However, our main interest is the bounds on the distribution of random variable $\pi^S(X^\ell)$.

The following lemma connects the distribution of random variable $\pi^S(X^\ell)$ with the distribution of random variable $\rho^S(X^{2\ell})$.

Basic Lemma. *The distributions of the random variables $\pi^S(X^\ell)$ and $\rho^S(X^{2\ell})$ are related in the following way:*

1. *For $\ell \geq 2/\varepsilon^2$ the inequality*

$$P \left\{ \pi^S(X^\ell) > \varepsilon \right\} \leq 2P \left\{ \rho^S(X^{2\ell}) > \frac{\varepsilon}{2} \right\} \quad (14.1)$$

is valid.

2. *The inequality*

$$P \left\{ \rho^S(X^{2\ell}) > \varepsilon \right\} \leq P \left\{ \pi^S(X^\ell) > \frac{\varepsilon}{2} \right\} - \left(P \left\{ \pi^S(X^\ell) > \frac{\varepsilon}{2} \right\} \right)^2 \quad (14.2)$$

is valid.

Proof By definition we have

$$P \left\{ \rho^S(X^{2\ell}) > \frac{\varepsilon}{2} \right\} = \int_{X(2\ell)} \theta \left[\rho^S(X^{2\ell}) - \frac{\varepsilon}{2} \right] dP(X^{2\ell}),$$

where

$$\theta(u) = \begin{cases} 1 & \text{if } u > 0, \\ 0 & \text{if } u \leq 0. \end{cases}$$

Taking into account that the space $X(2\ell)$ of samples of size 2ℓ is the direct product of two subspaces, namely $X_1(\ell)$ and $X_2(\ell)$ (two half-samples of size

ℓ), one can assert that for any measurable function $\phi(x_1, \dots, x_{2\ell})$ the equality

$$\int_{X^{(2\ell)}} \phi(x_1, \dots, x_{2\ell}) dP(X^{2\ell}) = \int_{X_1(\ell)} \left[\int_{X_2(\ell)} \phi(x_1, \dots, x_{2\ell}) dP(X_1^\ell) \right] dP(X_2^\ell)$$

is valid (Fubini's theorem).

Therefore

$$P \left\{ \rho^S(X^{2\ell}) > \frac{\varepsilon}{2} \right\} = \int_{X_1(\ell)} dP(X_1^\ell) \int_{X_2(\ell)} \theta \left[\rho^S(X^{2\ell}) - \frac{\varepsilon}{2} \right] dP(X_2^\ell)$$

(in the inner integral the first half-sample is fixed.) Denote by \mathcal{Q} the following event in the space $X_1(\ell)$

$$\{\pi^S(x_1, \dots, x_\ell) > \varepsilon\}.$$

Bounding the domain of integration, we obtain

$$P \left\{ \rho^S(X^{2\ell}) > \frac{\varepsilon}{2} \right\} \geq \int_{\mathcal{Q}} dP(X_1^\ell) \int_{X_2(\ell)} \theta \left[\rho^S(X^{2\ell}) - \frac{\varepsilon}{2} \right] dP(X_2^\ell). \quad (14.3)$$

We now bound the inner integral on the right-hand side of inequality which we denote by **I**. Here the sample x_1, \dots, x_ℓ is fixed and is such that

$$\pi^S(x_1, \dots, x_\ell) > \varepsilon.$$

Hence there exists an $A^* \in S$ such that

$$|P(A^*) - \nu(A^*; x_1, \dots, x_\ell)| > \varepsilon.$$

Let, for example,

$$\nu(A^*; x_1, \dots, x_\ell) < P(A^*) - \varepsilon$$

(the case $\nu(A^*; x_1, \dots, x_\ell) > P(A^*) + \varepsilon$ is considered analogously). Then in order that the conditions

$$|\nu(A^*; x_1, \dots, x_\ell) - \nu(A^*; x_{\ell+1}, \dots, x_{2\ell})| > \frac{\varepsilon}{2}$$

be satisfied, it is sufficient that the relation

$$\nu(A^*; x_{\ell+1}, \dots, x_{2\ell}) > P(A^*) - \frac{\varepsilon}{2}$$

be fulfilled from which we obtain

$$\begin{aligned} I &\geq \int_{X_2(\ell)} \theta \left[\nu(A^*; x_{\ell+1}, \dots, x_{2\ell}) - P(A^*) + \frac{\varepsilon}{2} \right] dP(X_2^\ell) \\ &= \sum_{k/\ell > P(A^*) - \frac{\varepsilon}{2}} C_\ell^k [P(A)]^k [1 - P(A)]^{\ell-k}. \end{aligned}$$

The last sum exceeds $1/2$ for $\ell > 2/\varepsilon^2$. Returning to (14.3) we obtain that for $\ell > 2/\varepsilon^2$

$$P \left\{ \rho^S(X^{2\ell}) > \frac{\varepsilon}{2} \right\} \geq \frac{1}{2} \int_Q dP(X^\ell) = \frac{1}{2} P \{ \pi^S(X^\ell) \}.$$

The first inequality has been proved.

To prove the second inequality, we note that the inequality

$$|\nu(A; x_1, \dots, x_\ell) - \nu(A; x_{\ell+1}, \dots, x_{2\ell})| > \varepsilon$$

implies validation of either the inequality

$$|\nu(A; x_1, \dots, x_\ell) - P(A)| > \frac{\varepsilon}{2}$$

or the inequality

$$|\nu(A; x_{\ell+1}, \dots, x_{2\ell}) - P(A)| > \frac{\varepsilon}{2}.$$

Taking into account that half-samples X_1^ℓ and X_2^ℓ are statistically independent, we obtain

$$\begin{aligned} & P \left\{ \sup_{A \in S} |\nu(A; x_1, \dots, x_\ell) - \nu(A; x_{\ell+1}, \dots, x_{2\ell})| > \varepsilon \right\} \\ & \leq 1 - \left(1 - P \left\{ \sup_{A \in S} |\nu(A; x_1, \dots, x_\ell) - P(A)| > \frac{\varepsilon}{2} \right\} \right) \\ & \quad \times \left(1 - P \left\{ \sup_{A \in S} |\nu(A; x_{\ell+1}, \dots, x_{2\ell}) - P(A)| > \frac{\varepsilon}{2} \right\} \right). \end{aligned}$$

From the last inequality comes

$$P \left\{ \rho^S(X^{2\ell}) > \varepsilon \right\} \leq P \left\{ \pi^S(X^\ell) > \frac{\varepsilon}{2} \right\} - \left(P \left\{ \pi^S(X^\ell) > \frac{\varepsilon}{2} \right\} \right)^2,$$

the second inequality of the lemma.

The lemma has been proved.

14.3 ENTROPY OF THE SET OF EVENTS

Let X be a set, let S be a system of its subsets, and let $X^\ell = x_1, \dots, x_\ell$ be a sequence of elements x of size ℓ . Each set $A \in S$ determines a subset X_A^ℓ of this set consisting of elements belonging to A . We say that A induces a subset X_A^ℓ on the set X^ℓ .

Denote by

$$N^S(x_1, \dots, x_\ell)$$

the number of different subsets X_A^ℓ induced by the sets $A \in S$.

It is clear that

$$1 \leq N^S(x_1, \dots, x_\ell) \leq 2^\ell.$$

Assume that this function is measurable and consider the function

$$H^S(\ell) = E \log_2 N^S(x_1, \dots, x_\ell).$$

It is clear as well that

$$0 \leq H^S(\ell) \leq e. \quad (14.4)$$

We call this function *entropy of the set of events S on the samples of size ℓ* .

This function has the property of semi-additivity

$$H^S(n+m) \leq H^S(n) + H^S(m). \quad (14.5)$$

To prove this, consider the sample

$$x_1, \dots, x_n, x_{n+1}, \dots, x_{n+m}.$$

Any subset of this set that was induced by the event $A \in S$ contains a subset of the set

$$x_1, \dots, x_n$$

that was induced by A and also contains the subset of the set

$$x_{n+1}, \dots, x_{n+m}$$

that was induced by this event.

Since the value $N^S(x_1, \dots, x_n, x_{n+1}, \dots, x_{n+m})$ does not exceed the number of pairs of subsets, where any pair contains one subset that was obtained from x_1, \dots, x_n and one subsequence that was obtained from x_{n+1}, \dots, x_{n+m} .

Therefore

$$N^S(x_1, \dots, x_n, \dots, x_{n+m}) \leq N^S(x_1, \dots, x_n)N^S(x_{n+1}, \dots, x_{n+m}).$$

From this inequality we obtain

$$\log_2 N^S(x_1, \dots, x_n, \dots, x_{n+m}) \leq \log_2 N^S(x_1, \dots, x_n) + \log_2 N^S(x_{n+1}, \dots, x_{n+m}). \quad (14.6)$$

Averaging this relation, we obtain (14.5).

Remark. Using inequality (14.5) repeatedly, one can derive

$$H^S(k\ell) \leq kH^S(\ell). \quad (14.7)$$

14.4 ASYMPTOTIC PROPERTIES OF THE ENTROPY

In this section we formulate and prove some asymptotic properties of the entropy of the set of events on the sample of the size ℓ . We shall use these properties for proving the necessary and sufficient conditions of uniform convergence of the frequencies to their probabilities.

Lemma 14.1. *The sequence $\frac{H^S(\ell)}{\ell}$, $\ell = 1, 2, \dots$, has a limit c , $0 \leq c \leq 1$, as $\ell \rightarrow \infty$:*

$$\lim_{\ell \rightarrow \infty} \frac{H^S(\ell)}{\ell} = c$$

The proof of this lemma repeats the proof of the analogous lemma in information theory for Shannon's entropy.

Proof. Since

$$0 \leq \frac{H^S(\ell)}{\ell} \leq 1,$$

there exists a lower bound

$$c = \liminf_{\ell \rightarrow \infty} \frac{H^S(\ell)}{\ell},$$

where $0 \leq c \leq 1$.

Then for any $\varepsilon > 0$ a value ℓ_0 can be found such that

$$\frac{H^S(\ell_0)}{\ell_0} < c + \varepsilon. \quad (14.8)$$

Note that any arbitrary ℓ can be rewritten in the form

$$\ell = n\ell_0 + k,$$

where $n \geq 0$ and $k < \ell_0$. Using properties (14.5) and (14.7), one can obtain a bound

$$\begin{aligned} H^S(\ell) &= H^S(n\ell_0 + k) \\ &\leq nH^S(\ell_0) + H^S(k) \leq nH^S(\ell_0) + k \end{aligned}$$

From this we derive

$$\frac{H^S(\ell)}{\ell} \leq \frac{nH^S(\ell_0) + k}{n\ell_0 + k} < \frac{nH^S(\ell_0) + \ell_0}{n\ell_0} = \frac{H^S(\ell_0)}{\ell_0} + \frac{1}{n}.$$

Using the condition (14.8), one obtains

$$\frac{H^S(\ell)}{\ell} < c + \varepsilon + \frac{1}{n}.$$

Since ℓ tends to infinity, n tends to infinity as well. We obtain

$$\limsup_{\ell \rightarrow \infty} \frac{H^S(\ell)}{\ell} < c + \varepsilon,$$

and because s is arbitrary we find that

$$\lim_{\ell \rightarrow \infty} \frac{H^S(\ell)}{\ell} = c.$$

The lemma has been proved.

Lemma 14.2. *For any ℓ the $\frac{H^S(\ell)}{\ell}$ is an upper bound for limit*

$$c = \lim_{\ell \rightarrow \infty} \frac{H^S(\ell)}{\ell},$$

in other words,

$$\frac{H^\Lambda(\ell)}{\ell} \geq \lim_{\ell \rightarrow \infty} \frac{H^\Lambda(\ell)}{\ell}.$$

Proof: The assertion of this lemma actually was proved in the proof of Lemma 14.1. In Lemma 14.1 we obtained the inequality

$$\frac{H^S(\ell)}{\ell} < \frac{H^S(\ell_0)}{\ell_0} + \frac{1}{n},$$

which is valid for arbitrary n , ℓ_0 and $\ell = n\ell_0$. As was proved in Lemma 14.1, the ratio $H^S(\ell)/\ell$ converges to some limit. Therefore for any ε there exist some n_0 such that for any $n > n_0$ the inequality

$$\lim_{n \rightarrow \infty} \frac{H^S(\ell)}{\ell} < \frac{H^S(\ell_0)}{\ell_0} + \frac{1}{n} + \varepsilon$$

holds. Since ε , ℓ_0 , and n are arbitrary values, we have

$$\lim_{\ell \rightarrow \infty} \frac{H^S(\ell)}{\ell} \leq \frac{H^S(\ell_0)}{\ell_0}$$

for any ℓ_0 .

The lemma is proved.

Now consider the random variable

$$r^S(x_1, \dots, x_\ell) = \frac{1}{\ell} \log_2 N^S(x_1, \dots, x_\ell).$$

The next lemma shows that when $\ell \rightarrow \infty$ this random variable converges (in probability) to the same limit to which the sequence $H^S(\ell)/\ell$ converges.

Lemma 143. *If the convergence*

$$\lim_{\ell \rightarrow \infty} \frac{H^S(\ell)}{\ell} = c,$$

takes place, then the random variable

$$r^S(x_1, \dots, x_\ell) = \frac{1}{\ell} \log_2 N^S(x_1, \dots, x_\ell)$$

converges in probability to c ; that is,

$$\lim_{\ell \rightarrow \infty} P\{|r^S(x_1, \dots, x_\ell) - c| > \varepsilon\} = 0.$$

In addition, the probabilities

$$P^+(\varepsilon, \ell) = P\{r^S(x_1, \dots, x_\ell) - c > \varepsilon\},$$

$$P^-(\varepsilon, \ell) = P\{c - r^S(x_1, \dots, x_\ell) > \varepsilon\}$$

satisfy the conditions

$$\sum_{\ell=1}^{\infty} P^+(\varepsilon, \ell) < \infty,$$

$$\lim_{\ell \rightarrow \infty} P^-(\varepsilon, \ell) = 0.$$

Proof. First, we estimate $P^+(\varepsilon, \ell)$. Since

$$\lim_{\ell \rightarrow \infty} \frac{H^S(\ell)}{\ell} = c,$$

for any $\varepsilon > 0$ there exists an ℓ_0 such that the inequality

$$\frac{H^S(\ell_0)}{\ell_0} < c + \frac{\varepsilon}{3}$$

holds true.

Consider the random sequence

$$q(n) = \frac{\sum_{i=0}^{n-1} \log_2 N^S(x_{i\ell_0+1}, \dots, x_{(i+1)\ell_0})}{\ell_0} = \sum_{i=0}^{n-1} r^\Lambda(x_{i\ell_0+1}, \dots, x_{(i+1)\ell_0}).$$

Note that the sequence $q(n)/n$, $n = 1, 2, \dots$, is an average of n random independent variables with expectation $H^S(\ell_0)/\ell_0$. Therefore the expectation of the random variable $q(n)/n$ equals to $H^\Lambda(\ell_0)/\ell_0$ as well.

Since the random variable r^Λ is bounded, it possesses the central moments of any order:

$$0 \leq r^S(x_1, \dots, x_{\ell_0}) \leq 1.$$

Let M_2 and M_4 be the central moments of order two and four. It is clear that

$$M_2 \leq 1, \quad M_4 \leq 1.$$

Then the central moment of order 4 for random variable $q(n)/n$ is

$$\frac{M_4}{n^3} + 3 \frac{n-1}{n^3} M_2^2 < \frac{1}{n^2}.$$

Using Chebyshev's inequality for moment of order 4, we obtain

$$P \left\{ \frac{q(n)}{n} - \frac{H^\Lambda(\ell_0)}{\ell_0} > \delta \right\} < \frac{4}{n^2 \delta}. \quad (14.9)$$

According to (14.6), the inequality

$$\frac{1}{n\ell_0} \log_2 N^S(x_1, \dots, x_{n\ell_0}) \leq \frac{1}{n\ell_0} \sum_{i=0}^{n-1} \log_2 N^S(x_{i\ell_0+1}, \dots, x_{(i+1)\ell_0})$$

holds true; that is,

$$r_{n\ell_0}^S = r^\Lambda(x_1, \dots, x_{n\ell_0}) \leq \frac{q(n)}{n}.$$

From (14.9) and the last inequality, we obtain

$$P \left\{ r_{n\ell_0}^S - \frac{H^S(\ell_0)}{\ell_0} > \delta \right\} < \frac{4}{n^2 \delta^4}.$$

Now let $\delta = \varepsilon/3$. Taking into account that

$$\frac{H^S(\ell_0)}{\ell_0} < c + \varepsilon,$$

we obtain the inequality

$$P \left\{ r_{n\ell_0}^S > c + \frac{2\varepsilon}{3} \right\} < \frac{244}{\varepsilon^4 n^2}. \quad (14.10)$$

For arbitrary $\ell > \ell_0$ we write $\ell = n\ell_0 + k$, where $n = [\ell/\ell_0]$ and $k < \ell_0$.

Since (14.6) we have

$$\log_2 N^S(x_1, \dots, x_\ell) \leq \log_2 N^S(x_1, \dots, x_{n\ell_0}) + k.$$

Therefore

$$r^S(x_1, \dots, x_\ell) \leq \frac{\log_2 N^S(x_1, \dots, x_{n\ell_0}) + k}{n\ell_0 + k}$$

Reinforcing this inequality, we obtain

$$r^S(x_1, \dots, x_\ell) \leq \frac{\log_2 N^S(x_1, \dots, x_{n\ell_0}) + \ell_0}{n\ell_0} = r^S(x_1, \dots, x_{\ell_0}) + \frac{1}{n}. \quad (14.11)$$

Now let ℓ be so large that the inequality

$$\frac{1}{n} < \frac{\varepsilon}{3}$$

holds. Then (14.10) and (14.11) imply the inequality

$$P^+(\varepsilon, \ell) = P \left\{ r_\ell^S > c + \varepsilon \right\} < \frac{244}{\varepsilon^4 n^2}. \quad (14.12)$$

Note that $n \rightarrow \infty$ when $\ell \rightarrow \infty$. Taking this remark into account, we conclude from (14.12) that the inequality

$$\lim_{\ell \rightarrow \infty} P^+(\varepsilon, \ell) = 0$$

holds true.

Besides, from the same reason the inequality

$$\sum_{\ell=1}^{\infty} P^+(\varepsilon, \ell) < \infty \quad (14.13)$$

holds true.

Indeed, since $P^+(\varepsilon, 4) \leq 1$ it is sufficient to evaluate the part of this sum starting with some large ℓ_0 . For this part of the sum we obtain the bound using inequality (14.12) inequality. Therefore we have

$$\sum_{\ell=1}^{\infty} P^+(\varepsilon, \ell) < \ell_0 + \sum_{n=\ell_0}^{\infty} \frac{244}{\varepsilon^4 n^2} < \infty.$$

Thus the first part of the lemma is proved.

To prove the lemma, it remains to show that the equality

$$\lim_{\ell \rightarrow \infty} P^-(\varepsilon, \ell) = 0$$

holds true.

Consider ℓ_0 such that for all $\ell > \ell_0$ the inequality

$$\left| \frac{H^S(\ell)}{\ell} - c \right| < \frac{\varepsilon}{2}$$

holds true. Note that $H^S(\ell)/\ell$ is the expectation of the random variable $r_\ell^S = r^S(x_1, \dots, x_\ell)$. Let us write this fact in the form

$$\int_{r=0}^{H^S(\ell)/\ell} \left(\frac{H^S(\ell)}{\ell} - r_\ell^S \right) dP(r^S) = \int_{H^S(\ell)/\ell}^{r=1} \left(r_\ell^S - \frac{H^S(\ell)}{\ell} \right) dP(r^S).$$

Denote the right-hand side of this equality with R_1 and denote left-hand side of the equality with R_2 . When $\ell > \ell_0$ we have

$$R_1 \geq \frac{\varepsilon}{2} \int_{r_\ell^S=0}^{r_\ell^S=c-\varepsilon} dP(r_\ell^S) = \frac{\varepsilon}{2} P^-(\varepsilon, Y).$$

Now let $\delta > 0$ be an arbitrary small value. Then

$$\begin{aligned} R_2 &= \int_{r_\ell^S=H^S(\ell)/\ell}^{r_\ell^S=c+\delta} \left(r_\ell^S - \frac{H^S(\ell)}{\ell} \right) dP(r_\ell^S) \\ &\quad + \int_{r_\ell^S=c+\delta}^{r_\ell^S=1} \left(r_\ell^S - \frac{H^S(\ell)}{\ell} \right) dP(r_\ell^S) \\ &\leq \left| c + \delta - \frac{H^S(\ell)}{\ell} \right| + \int_{r_\ell^S=c+\delta}^{r_\ell^S=1} dP(r_\ell^S). \end{aligned}$$

Using our notation, we rewrite this inequality in the form

$$R_2 \leq \left| c + \delta - \frac{H^S(\ell)}{\ell} \right| + P^+(\delta, \ell).$$

Combining the estimates for R_1 and R_2 , we have

$$\frac{\varepsilon}{2} P^-(\varepsilon, \ell) \leq \left| c + \delta - \frac{H^A(\ell)}{\ell} \right| + P^+(\delta, \ell).$$

In the case when ℓ tends to infinity, we obtain

$$\lim_{\ell \rightarrow \infty} P^-(\varepsilon, \ell) \leq \frac{26}{\varepsilon}.$$

Since δ is an arbitrary small value and $P_-(s, \ell)$ is a positive value, we conclude that

$$\lim_{\ell \rightarrow \infty} P^-(\varepsilon, \ell) = 0$$

The lemma is proved.

14.5 NECESSARY AND SUFFICIENT CONDITIONS OF UNIFORM CONVERGENCE. PROOF OF SUFFICIENCY

Chapter 3 formulated the theorem according to which the convergence

$$\lim_{\ell \rightarrow \infty} \frac{H^S(\ell)}{\ell} = 0$$

is the necessary and sufficient condition for uniform convergence (in probability) of the frequencies to their probabilities over a given set of events. In this section we will prove a stronger assertion.

Theorem 14.1. *Let the functions $N^S(x_1, \dots, x_\ell)$, $\pi^S(x_1, \dots, x_\ell)$, $\rho^S(x_1, \dots, x_\ell)$ be measurable for all ℓ .*

Then:

If the equality

$$\lim_{\ell \rightarrow \infty} \frac{H^S(\ell)}{\ell} = 0, \quad (14.14)$$

holds true, then the uniform convergence takes place with probability one (almost surely).

If, however, the equality

$$\lim_{\ell \rightarrow \infty} \frac{H^S(\ell)}{\ell} = c > 0, \quad (14.15)$$

holds true, then there exists $\delta(c) > 0$ which does not depend on ℓ such that

$$\lim_{\ell \rightarrow \infty} P\{\pi^S(x_1, \dots, x_\ell) > 6\} = 1;$$

that is, the probability that a maximal (over given set of events) deviation of the frequency from the corresponding probability exceeds 6 tends to one with increasing number of ℓ .

Therefore from this theorem we see that equality (14.14) is the necessary and sufficient condition for uniform convergence of the frequencies to their probabilities over a given set of functions. In this section we prove the first part of this theorem: sufficiency of condition (14.14) for uniform convergence with probability one. The second part of this theorem will be proven in the next two sections of this chapter.

The proof of this part of the theorem actually repeats the proof of the Theorem 4.1.

Proof of the Sufficiency of the Condition (14.14) for Uniform Convergence with Probability One. Suppose that the equality

$$\lim_{\ell \rightarrow \infty} \frac{H^\Lambda(\ell)}{\ell} = 0$$

holds true. Let us evaluate the value

$$P \left\{ \sup_{A \in S} |\nu(A; x_1, \dots, x_\ell) - P(A)| > \varepsilon \right\} = P \left\{ \pi_\ell^S > \varepsilon \right\}.$$

According to the Basic Lemma, the inequality

$$P \left\{ \pi_\ell^S > \varepsilon \right\} < 2P \left\{ \rho_\ell^S > \frac{\varepsilon}{2} \right\}$$

holds true.

On the other hand, it was shown in the proof of the Theorem 4.1 that the equality

$$P \left\{ \rho_\ell^S > \frac{\varepsilon}{2} \right\} = \frac{1}{(2l)!} \int_{X^{(2l)}} \sum_{i=1}^{(2l)!} \theta \left[\rho^S(T_i X^{2l}) - \frac{\varepsilon}{2} \right] dP(X^{2l})$$

holds true, where T_i are all possible permutations of the sequence x_1, \dots, x_{2l} . Besides, in Chapter 4, Section 4.13 it was shown that

$$\begin{aligned} K &= \frac{1}{(2l)!} \sum_{i=1}^{(2l)!} \theta \left[\rho^S(T_i X^{2l}) - \frac{\varepsilon}{2} \right] \\ &< 3N^S(x_1, \dots, x_{2l}) \exp \frac{-\varepsilon^2 l}{4} \end{aligned}$$

Note that for sufficiently large ℓ the value K does not exceed 1.

Now, let us divide the region of integration into two parts: subregion X_1 , where

$$\frac{\log_2 N^S(x_1, \dots, x_{2l})}{2l} \leq \frac{\varepsilon^2}{8},$$

and subregion X_2 , where

$$\frac{\log N^S(x_1, \dots, x_{2\ell})}{2f?} > \frac{\varepsilon^2}{8}.$$

Then using a majorant for K , we obtain

$$P \left\{ \rho_{2\ell}^S > \frac{\varepsilon}{2} \right\} < 2 \int_{X_1} N^S(x_1, \dots, x_{2\ell}) e^{-\varepsilon^2 \ell/4} dP(X^{2\ell}) + \int_{X_2} dP(X^{2\ell}).$$

Note that since

$$\lim_{\ell \rightarrow \infty} \frac{H^S(\ell)}{\ell} = 0,$$

we have

$$\int_{X_2} dP(X^{2\ell}) = P^+ \left(\frac{\varepsilon^2}{8}, 2\ell \right)$$

(for $P^+ \left(\frac{\varepsilon^2}{8}, 2\ell \right)$ see Section 14.4). Taking into account that in the region X_1 the inequality

$$N^S(x_1, \dots, x_{2\ell}) \leq 2^{\varepsilon^2 \ell/4}$$

holds true, we obtain

$$P \left\{ \rho_{2\ell}^S > \frac{\varepsilon}{2} \right\} < 2 \cdot 2^{\varepsilon^2 \ell/4} \cdot e^{-\varepsilon^2 \ell/4} + P^+ \left(\frac{\varepsilon^2}{8}, 2\ell \right). \quad (14.16)$$

The first term on the right-hand side of inequality (14.16) tends to zero when $\ell \rightarrow \infty$, and the second term of the inequality tends to zero in accordance with Lemma 14.3. Even more, since in accordance with this lemma the inequality

$$\sum_{\ell=1}^{\infty} P^+ \left(\frac{\varepsilon^2}{8}, \ell \right) < \infty$$

is valid, then the inequality

$$\sum_{\ell=1}^{\infty} P \left\{ \rho^S(x_1, \dots, x_{2\ell}) > \frac{\varepsilon}{2} \right\} < \infty$$

is valid as well. The last inequality implies the inequality

$$\sum_{\ell=1}^{\infty} P \{ \pi^S(x_1, \dots, x_{2\ell}) > \varepsilon \} < \infty.$$

According to the Borel–Cantelli lemma, this inequality implies the convergence of frequencies to their probabilities with probability one.

Thus, the first part of the theorem has been proven.

14.6 NECESSARY AND SUFFICIENT CONDITIONS OF UNIFORM CONVERGENCE. PROOF OF NECESSITY

Now let the equality

$$\lim_{\ell \rightarrow \infty} \frac{H^S(\ell)}{\ell} = c > 0$$

be valid. In accordance with the Basic Lemma, if the equality

$$\lim_{\ell \rightarrow \infty} P\{\rho^S(x_1, \dots, x_{2\ell}) > 2\delta\} = 1 \quad (14.17)$$

is valid, then the equality

$$\lim_{\ell \rightarrow \infty} P\{\pi^S(x_1, \dots, x_{2\ell}) > \delta\} = 1$$

is also valid.

Therefore to prove the second part of the theorem it is sufficient to show that under condition (14.15), equality (14.17) holds true for some $\delta = \delta(c)$.

To clarify the idea of proving this part of the theorem, let us consider its particular case, namely the case when equality

$$\lim_{\ell \rightarrow \infty} \frac{H^S(\ell)}{\ell} = 1$$

holds true. In this case, as was shown in the remark to Lemma 14.2, the equality

$$\frac{H^S(\ell)}{\ell} = 1$$

is valid for any ℓ .

Since $H^S(\ell)/\ell$ is the mathematical expectation of the random variable

$$\frac{\log_2 N^S(x_1, \dots, x_\ell)}{\ell} \leq 1$$

the equality

$$P\left\{\frac{\log_2 N^S(x_1, \dots, x_\ell)}{\ell} = 1\right\} = 1$$

is valid. This means that for any finite ℓ with probability one, the equality

$$N^S(x_1, \dots, x_\ell) = 2^\ell$$

is valid; that is, almost any sample x_1, \dots, x_ℓ induced all 2^ℓ possible subsets by events of the set S . In particular, this means that for almost any sample $x_1, \dots, x_{2\ell}$ an event $A^* \in S$ can be found such that

$$x_i \in A^*, \quad i = 1, 2, \dots, \ell,$$

$$x_i \notin A^*, \quad i = \ell + 1, \dots, 2\ell$$

Then

$$\nu_1(A^*; x_1, \dots, x_\ell) = 1,$$

$$\nu_2(A^*; x_{\ell+1}, \dots, x_{2\ell}) = 0$$

and therefore with probability one we obtain

$$\sup_{A \in S} |\nu_1(A^*; x_1, \dots, x_\ell) - \nu_2(A^*; x_{\ell+1}, \dots, x_{2\ell})| = 1.$$

In this case for any $\delta \leq 0.5$ the equality

$$\lim_{\ell \rightarrow \infty} P \left\{ \sup_{A \in S} |\nu_1(A^*; x_1, \dots, x_\ell) - \nu_2(A^*; x_{\ell+1}, \dots, x_{2\ell})| \geq 2\delta \right\} = 1$$

is valid.

The idea of proving the necessity of (14.15) in the general case is based on the fact that if the equality

$$\lim_{\ell \rightarrow \infty} \frac{H^S(\ell)}{\ell} = c > 0$$

holds, then from almost any sample of size ℓ the subsample of size $n(\ell)$ can be subtracted where $n(\ell)$ is a monotonously increasing function such that this subsample can be shattered by events from the $A \in S$ in all $2^{n(\ell)}$ possible ways.

To implement this idea we need the following lemma.

Lemma 14.4. *Suppose that for some a ($0 < a \leq 1$), some $\ell > 9/a^2$, and some sample*

$$x_1, \dots, x_\ell$$

the inequality

$$N^S(x_1, \dots, x_\ell) \geq 2^{a\ell}$$

holds true.

Then the subsample

$$x_{i_1}, \dots, x_{i_r}$$

of size

$$r = [ql],$$

where $q = a^2 e / 9$ (e – is the basis of logarithm), can be found such that the equality

$$N^S(x_{i_1}, \dots, x_{i_r}) = 2$$

holds true.

Proof: According to Lemma 4.2 (for this lemma see Chapter 4, Section 4.3), this subsample exists if the inequality

$$N^S(x_1, \dots, x_\ell) > \sum_{i=1}^{r-1} C_\ell^i = \Phi(\ell, r)$$

is valid.

To prove this lemma it is sufficient to check the inequality

$$2^{a\ell} > \Phi(\ell, r). \quad (14.18)$$

Taken into account that for the case $r \geq 2$ and $\ell \geq r + 1$ one can use the bound for function $\Phi(\ell, r)$ obtained in Chapter 4, Section 4.10, we obtain

$$\Phi(\ell, r) < \left(\frac{\ell e}{r} \right)^r.$$

Note that the function $(\ell e/t)^t$ monotonously increases with t when $t < 1$. Therefore

$$\Phi(\ell, r) < \left(\frac{e}{q} \right)^{q\ell},$$

where

$$r = [ql] \leq ql.$$

The relationship (14.18) will be proved if we prove the inequality

$$2^{a\ell} > \left(\frac{e}{q} \right)^{q\ell}.$$

Taking the logarithm of both sides of this inequality and simplifying the expression, we obtain

$$a > q \log_2 \left(\frac{e}{q} \right). \quad (14.19)$$

Note that for $z > 0$ the following inequality

$$\log_2 z \leq \frac{2 \log e}{e} \sqrt{z}$$

holds true.

Indeed, this inequality is true because the function $\log_2 z / \sqrt{z}$ achieves its maximum at the point $z = e^2$ and the maximum of this function equals $2 \log e / e$. Therefore the inequality

$$a > \sqrt{eq} \frac{2 \log e}{e}$$

implies inequality (14.19). When $q = a^2 e / 9$, this inequality holds true since the equality

$$a > \frac{2 \log e}{3} a$$

is true.

The lemma has been proved.

Recall that according to Lemma 14.3 if

$$\lim_{\ell \rightarrow \infty} \frac{H^S(\ell)}{\ell} = c > 0,$$

then the probability

$$P \left\{ \frac{\lg N^S(x_1, \dots, x_\ell)}{\ell} > c - \delta \right\}$$

tends to one when ℓ tends to infinity and $\delta > 0$. Hence for sufficiently large ℓ with probability arbitrarily close to one, the inequality

$$N^S(x_1, \dots, x_\ell) > 2^{(c/2)\ell} \quad (14.20)$$

holds true. According to Lemma 14.4, in this case from any sample the sub-sample of the size

$$r = \left[q \left(\frac{c}{2} \ell \right) \right]$$

can be found such that it induces all possible 2^r subsets by the set of events S .

This fact gives a key to prove the necessary condition of the theorem.

Scheme of the Proving the Necessity of the Conditions of the Theorem. To prove the necessity of the conditions of the theorem we have to show that for some $\delta(c)$ the equality

$$\lim_{\ell \rightarrow \infty} P\{\rho^S(x_1, \dots, x_{2\ell}) > 2\delta\} = 1$$

is valid. To prove this fact we compare the frequencies of occurrence of the events on the first half-sample and on the second half-sample. To do this we take a sample of size 2ℓ and then split it randomly into two subsamples of equal size. Then for any events of the set S we calculate and compare the number of occurred events of these subsets.

Now let us consider another scheme. Suppose that the sample of size \mathbf{P} satisfies the condition

$$N^S(x_1, \dots, x_{2\ell}) > 2^{c\ell}.$$

Then one can extract from the $x_1, \dots, x_{2\ell}$ the sample X' of size

$$r = \left[2q \left(\frac{c}{2} \right) \ell \right],$$

on which all subsamples can be induced.

Now let us randomly split this sample into two equal subsamples: subsamples $X_1^{r/2}$ and subsample $X_2^{r/2}$. Then let us independently split the remainder $X^{2\ell}/X'$ into two equal subsamples: subsample $X_1^{\ell-r/2}$ and subsample $X_2^{\ell-r/2}$. According to the construction there exists such event \mathbf{A}^* that all elements of $X_1^{r/2}$ belong to \mathbf{A}^* and all elements of $X_2^{r/2}$ do not belong to \mathbf{A}^* . Suppose that in the subsamples $X_1^{\ell-r/2}$ and $X_2^{\ell-r/2}$ there are m elements that belong to \mathbf{A}^* . Approximately half of them belong to $X_1^{\ell-r/2}$, while the other half belong to $X_2^{\ell-r/2}$. Then

$$|\nu_1(A^*) - \nu_2(A^*)| \sim \frac{r}{2\ell} + \frac{m}{2\ell} - \frac{m}{2\ell} = \frac{r}{2\ell} \sim q$$

and consequently

$$\sup_{A \in S} |\nu_1(A^*) - \nu_2(A^*)| > q.$$

Because $q > 0$ does not depend on the length of the sample, there is no uniform convergence.

Of course this scheme is not equivalent to the initial one since the sample X' and the remainder $X^{2\ell}/X'$ are not necessarily split into two equal parts when one splits sample $X^{2\ell}$ into two equal parts. However, for sufficiently large ℓ (and consequently r), these conditions are fulfilled rather precisely.

In the next section we will give the formal proof, which takes into account all these assumptions and approximations.

14.7 NECESSARY AND SUFFICIENT CONDITIONS. CONTINUATION OF PROVING NECESSITY

Let the equality

$$\lim_{\ell \rightarrow \infty} \frac{H^S(\ell)}{\ell} = c > 0$$

hold true.

To prove the necessity, we just have to estimate the quantity

$$P\{\rho^S(x_1, \dots, x_{2\ell}) > 2\delta\} = \frac{1}{(2\ell)!} \int_{X^{(2\ell)}} \sum_i [\theta(T_i X^{2\ell}) - 2\delta] dP(X^{2\ell}),$$

where T_i , $i = 1, \dots, (2\ell)!$, are all $(2\ell)!$ possible permutations of the sample $x_1, \dots, x_{2\ell}$. Denote by $K(X^{2\ell})$ the integrand and reduce the domain of integration

$$P\{\rho^S(x_1, \dots, x_{2\ell}) > 2\delta\} \geq \frac{1}{(2\ell)!} \int_{\log_2 N^S(X^{2\ell}) > \frac{c}{2}} K(X^{2\ell}) dP(X^{2\ell}).$$

Now let us examine the integrand $K(X^{2\ell})$ assuming that

$$\frac{\log_2 N^S(X^{2\ell})}{2\ell} > \frac{c}{2};$$

that is,

$$N^S(X^{2\ell}) > 2^{c\ell}.$$

Let us choose

$$0 < q(c) < \frac{1}{2}$$

in such a way that (in accordance with Lemma 14.4) for sufficiently large ℓ the subsample X^n of the size $n > q\ell$ exists in which the set of events S induces all possible subsamples (i.e., $N^S(X^n) = 2^n$). Now we choose $\delta(c) = q/8$. Note that values q and 6 do not depend on ℓ .

Observe that all permutations T_i can be classified into the groups R_s corresponding to some partition of the sample $x_1, \dots, x_{2\ell}$ into first and second half-samples.

It is clear that the quantity

$$\rho(T_i X^{2\ell}) = \sup_{A \in S} |\nu_1(A; T_i X^{2\ell}) - \nu_2(A; T_i X^{2\ell})|$$

depends only on the group R_s and does not depend on specific transmutation T_i into the group. Therefore

$$K = \frac{1}{C_{2\ell}^\ell} \sum_s \theta [\rho^S(R_s X^{2\ell}) - 2\delta],$$

where the summation is taken over all different separations the sample into two subsamples.

Now, let X^n be the subsample of the size n in which the set of events S induces all 2^n subsamples. Denote by $X^{2\ell-n}$ the complement of X^n with respect to the $X^{2\ell}$ (the number of elements in the $X^{2\ell-n}$ equals $2\ell - n$).

The partition R_s of the sample $X^{2\ell}$ into two half-samples is completely described if the partition R_k^1 of the subsample X^n into the two subsamples and the partition R_j^2 of the sample $X^{2\ell-n}$ into the two subsamples are given.

Let $R_s = R_k^1 R_j^2$. Let $r(k)$ be the number of elements in the subsample X^n which belong under partition R_k^1 to the first half-sample and let $m(j)$ be the number of elements of the subsample $X^{2\ell-n}$ which belong under partition R_j^2 to the first half-sample. Clearly, $r(k) + m(j) = \ell$ for k and j corresponding to the same partition R_s . We have

$$\bar{K} = \frac{1}{C_{2\ell}^\ell} \sum_k \sum_j^* \theta [\rho^S(R_k^1 R_j^2 X^{2\ell}) - 2\delta],$$

where \sum_j^* is the summation over only those j for which $m(j) = \ell - r(k)$, and

$$K = \frac{1}{C_{2\ell}^\ell} \sum_{r=0}^{\ell} \left(\sum_k^* \sum_j^* \theta [\rho^S(R_k^1 R_j^2 X^{2\ell}) - 2\delta] \right),$$

where \sum_k^* is summation over only those k for which $r(k) = r$. For each R_k^1 we can specify a set $A(k) \in S$ such that $A(k)$ includes exactly the elements of subsample X^n which belong under partition R_k^1 to the first half-sample.

Introduce the notations:

$p(k)$ is the number of the elements in the subsample $X^{2\ell-n}$ belonging to $A(k)$.

$t(k, j)$ is the number of the elements in $X^{2\ell-n}$ belonging, under partition R_j^2 , to the first half-sample.

Then

$$\nu_1(A(k); X^{2\ell}) = \frac{(r+t)}{\ell},$$

$$\nu_2(A(k); X^{2\ell}) = \frac{(p-t)}{\ell},$$

$$\rho(A(k); X^{2\ell}) = |\nu_1(A(k); X^{2\ell}) - \nu_2(A(k); X^{2\ell})| = \frac{|r+2t-p|}{\ell}$$

We further take into account that

$$\rho^S(X^{2\ell}) = \sup_{A \in S} \rho(A; X^{2\ell}) > \rho(A(k); X^{2\ell}).$$

Replacing $\rho^S(X^{2\ell})$ by $\rho(A(k); X^{2\ell})$ we estimate K to obtain

$$K \geq \frac{1}{C_{2\ell}^\ell} \sum_{r=0}^{\ell} \sum_k^* \left(\sum_j^* \theta \left[\frac{|2t(k, j) + r - p(k)|}{\ell} - 2\delta \right] \right).$$

Observe that the number of partitions R_j^1 satisfying the conditions $s(j) = \ell - r$ for fixed r is

$$C_{2\ell-r-p(k)}^{\ell-r},$$

and the number of partitions R_j^2 which in addition correspond to the same t for fixed r and $A(k)$ is

$$C_{p(k)}^t C_{2\ell-r-p(k)}^{\ell-r-t}.$$

Then the estimate for K is

$$K \geq \frac{1}{C_{2\ell}^\ell} \sum_r \sum_k \sum_t C_{p(k)}^t C_{2\ell-r-p(k)}^{\ell-r-t}.$$

After an elementary transformation, one obtains

$$K \geq \sum_r \frac{C_n^r C_{2\ell-n}^{\ell-r}}{C_{2\ell}^\ell} \sum_k \frac{1}{C_n^r} \sum_t \frac{C_{p(k)}^t C_{2\ell-n-p(k)}^{\ell-r-t}}{C_{2\ell-n}^{\ell-r}}, \quad (14.21)$$

where the summation on t are carried out in the limits determined by the expression

$$\frac{|r+2t-p|}{e} > 26. \quad (14.22)$$

Now let

$$0 < \varepsilon < \frac{q}{20}.$$

Consider the quantity K^* :

$$K^* = \sum_r \frac{C_n^r C_{2\ell-n}^{\ell-r}}{C_{2\ell}^\ell} \sum_k \frac{1}{C_n^r} \sum_t \frac{C_{p(k)}^t C_{2\ell-n-p(k)}^{\ell-r-t}}{C_{2\ell-n}^{\ell-r}},$$

which differs from (14.21) only in the limits of summation

$$|r - \frac{n}{2}| \leq \varepsilon n, \quad (14.23)$$

$$|t - \frac{p(k)(\ell-r)}{2\ell-n}| < \varepsilon \ell. \quad (14.24)$$

Observe that if r and t satisfy inequalities (14.23) and (14.24), then inequality (14.22) holds true. Taking into account these inequalities, we obtain

$$\begin{aligned}\frac{r+2t-p}{\ell} &> \frac{1}{\ell} \left[\frac{n}{2} - \varepsilon n + \frac{2p}{2\ell-n} \left(\ell - \frac{n}{2} - \varepsilon n \right) - 2\varepsilon\ell - p \right] \\ &\geq \frac{1}{\ell} \left[\frac{n}{2} - \varepsilon \left((n+2\ell + \frac{2pn}{2\ell-n}) \right) \right]\end{aligned}$$

As far as

$$q\ell < n < \ell, \quad \delta = \frac{q}{8}, \quad \varepsilon < \frac{q}{20}$$

from the last expression we obtain

$$\frac{r+2t-p}{\ell} > \frac{1}{\ell} \left[\frac{n}{2} - 5\varepsilon\ell \right] = \frac{q}{4} = 2\delta.$$

Since the domain of summation of K includes the domain of summation K^* , we have

$$K \geq K^*$$

Note that for any $\eta > 0$ there exists $\ell_0 = \ell_0(\eta, q)$ such that for all $\ell > \ell_0$ we have the inequality

$$\sum_r \frac{C_n^r C_{2\ell-n}^{\ell-r}}{C_{2\ell}^\ell} > 1 - \eta$$

(here the summation is taken over r which satisfy (14.23)) and the inequality

$$\sum_r \frac{C_p^t C_{2\ell-n-p}^{\ell-r-t}}{C_{2\ell-n}^{\ell-r}} > 1 - \eta \tag{14.25}$$

(here the summation is taken over t which satisfy (14.24)).

Indeed,

$$\frac{C_n^r C_{2\ell-n}^{\ell-r}}{C_{2\ell}^\ell}$$

is the probability to draw r black balls from the urn containing n black balls and $2\ell - n$ white balls, when one randomly draws ℓ balls without replacement. In this case the expectation of number of black balls in the sample is equal to $n/2$, and the right-hand side of inequality (14.25) is the probability that the deviation of the number of black balls from the expectation

of the black balls exceed εn . Since for the scheme of drawn balls without replacement the law of large numbers holds true starting for some large ℓ , the inequality (14.25) is valid.

Analogously, the quantity

$$\frac{C_p^t C_{2\ell-n-p}^{\ell-r-t}}{C_{2\ell-n}^{\ell-r}}$$

is the probability to draw t black balls from the urn containing p black balls and $2\ell - n - p$ white balls when one randomly draws $\ell - r$ balls without replacement. The expectation of the number of black balls in the sample is equal to

$$\frac{p(\ell - r)}{2\ell - n},$$

and consequently inequality (14.26) expresses the law of large numbers in this case.

Then, taken into account that the number of partitions R_k of the subsample X^n for fixed r is equal to C_ℓ^r , we obtain for $\ell > \ell_0$

$$K \geq (1 - \eta)^2$$

Thus, for $\ell > \ell_0$ and $6 = q/8$ we obtain

$$\begin{aligned} P\{\rho^S(X^{2\ell}) > 2\delta\} &\geq \int_{\frac{\log_2 N^S(X^{2\ell})}{\ell} > \frac{c}{2}} K(X^{2\ell}) dP(X^{2\ell}) \\ &\geq (1 - \eta)^2 \left(1 - P^-\left(\frac{c}{2}, \ell\right)\right). \end{aligned}$$

Since according to Lemma 14.2 we have

$$\lim_{\ell \rightarrow \infty} P^-\left(\frac{c}{2}, \ell\right) = 0,$$

we obtain

$$\lim_{\ell \rightarrow \infty} P\{\rho^S(X^{2\ell}) > 2\delta\} \geq (1 - \eta)^2.$$

Taken into account that η is arbitrarily small, we conclude that the equality

$$\lim_{\ell \rightarrow \infty} P\{\rho^S(X^{2\ell}) > 2\delta\} = 1$$

holds true.

Thus the theorem is proved.

15

NECESSARY AND SUFFICIENT CONDITIONS FOR UNIFORM CONVERGENCE OF MEANS TO THEIR EXPECTATIONS

In Chapter 14 we obtained the necessary and sufficient conditions for uniform (two-sided) convergence of frequencies to their probabilities over a given set of events that also can be described in terms of uniform two-sided convergence of the means to their expectations for a given set of indicator functions.

In this chapter we will generalize these results to the set of bounded real-valued functions

$$a \leq F(x, \alpha) \leq b. \quad (15.1)$$

Below, without loss of generality we assume that $a = 0$ and $b = 1$. (Note that indicator functions satisfy the conditions.) We are looking for the necessary and sufficient conditions of uniform convergence of the means to their expectations over a given set of functions $F(x, \alpha)$, $\alpha \in A$; in other words, we are looking for conditions under which the limit

$$\lim_{\ell \rightarrow \infty} P \left\{ \sup_{\alpha \in A} \left| EF(x, \alpha) - \frac{1}{\ell} \sum_{i=1}^{\ell} F(x_i, \alpha) \right| > \varepsilon \right\} = 0 \quad (15.2)$$

holds true.

15.1 ε ENTROPY

We start with some definitions.

Let A be a bounded set of vectors in R^ℓ . Consider a finite set of $T \subset R^\ell$

such that for any $y \in A$ there exists an element $t \in T$ satisfying

$$\rho(t, y) < \varepsilon.$$

We call this set T a *relative ε net* of $A \in R^\ell$.

Below we shall assume that the metric is defined by

$$\rho(t, y) = \max_{1 \leq i \leq n} |t^i - y^i|, \quad t = (t^1, \dots, t^n), \quad y = (y^1, \dots, y^n),$$

and the norm of a vector z is given by

$$\|z\| = \max_{1 \leq i \leq n} |z^i|.$$

If an ε net T of a set A is such that $T \subset A$, then we call it a proper ε net of the set A .

The minimal number of elements in an ε net of the set A relative to R^ℓ will be denoted by $N(\varepsilon, A)$, and the minimal number of elements in a proper ε net is denoted by $N_0(\varepsilon, A)$. It is easy to see that

$$N_0(\varepsilon, A) \geq N(\varepsilon, A). \quad (15.3)$$

On the other hand,

$$N_0(2\varepsilon, A) < N(\varepsilon, A). \quad (15.4)$$

Indeed, let T be a minimal ε net of A relative to R^ℓ . We assign to each element $t \in T$ an element $y \in A$ such that $\rho(t, y) < \varepsilon$ (such an element y always exists, since otherwise the ε net could have been reduced). The totality T_0 of elements of this kind forms a proper 2ε net in A (for each $y \in A$ there exists $t \in T$ such that $\rho(t, y) < \varepsilon$, and for such a $t \in T$ there exists $\tau \in T_0$ such that $\rho(\tau, t) < \varepsilon$ and hence $\rho(y, \tau) < 2\varepsilon$).

Let $F(x, \alpha)$ be a class of real functions in the variable $x \in X$ depending on an abstract parameter $\alpha \in A$. Let

$$x_1, \dots, x_\ell$$

be a sample. Consider in the space R^ℓ a set A of vectors z with coordinates

$$z^i = F(x_i, \alpha), \quad i = 1, \dots, \ell$$

formed by all $\alpha \in A$.

If the condition $0 \leq F(x, \alpha) \leq 1$ is fulfilled, then the set

$$A = A(x_1, \dots, x_\ell)$$

belongs to an ℓ -dimensional cube $0 \leq z^i \leq 1$ and is therefore bounded and possesses a finite ε net.

The number of elements of a minimal relative ε net of A in R^ℓ is

$$N(\varepsilon; A(x_1, \dots, x_\ell)) = N^\Lambda(x_1, \dots, x_\ell; \varepsilon)$$

and the number of elements of a minimal proper ε net is $N_0^\Lambda(x_1, \dots, x_\ell; \varepsilon)$. If a probability measure $P(x)$ is defined on X and x_1, \dots, x_ℓ is an independent random sample and $N^\Lambda(x_1, \dots, x_\ell; \varepsilon)$ is a function measurable with respect to this measure on sequences x_1, \dots, x_ℓ , then there exists an expected entropy (or simply an ε entropy)

$$H^\Lambda(\varepsilon, \ell) = E \ln N^\Lambda(x_1, \dots, x_\ell; \varepsilon).$$

It is easy to verify that a minimal relative ε net satisfies

$$N^\Lambda(x_1, \dots, x_{\ell+k}; \varepsilon) \leq N^\Lambda(x_1, \dots, x_\ell; \varepsilon) N^\Lambda(x_{\ell+1}, \dots, x_{\ell+k}; \varepsilon). \quad (15.5)$$

(Recall that $\rho(z_1, z_2) = \max_{1 \leq i \leq n} |z_1^i - z_2^i|$.)

Indeed, in this case a direct product of relative ε nets is also a relative ε net. Thus,

$$H^\Lambda(\varepsilon, \ell+k) \leq H^\Lambda(\varepsilon, \ell) + H^\Lambda(\varepsilon, k). \quad (15.6)$$

At the end of this section it will be shown that there exists the limit

$$c(\varepsilon) = \lim_{\ell \rightarrow \infty} \frac{H^\Lambda(\varepsilon, \ell)}{\ell} \quad 0 \leq c(\varepsilon) \leq \ln \left[1 + \frac{1}{\varepsilon} \right]$$

and that the convergence

$$\frac{\ln N^\Lambda(x_1, \dots, x_\ell; \varepsilon)}{\ell} \xrightarrow[\ell \rightarrow \infty]{P} c(\varepsilon) \quad (15.7)$$

holds.

We will consider two cases:

1. The case where for all $\varepsilon > 0$ the following equality holds true:

$$\lim \frac{H^\Lambda(\varepsilon, \ell)}{\ell} = c(\varepsilon) = 0.$$

2. The case where there exists an ε_0 such that we have $c(\varepsilon_0) > 0$ (then also for all $\varepsilon < \varepsilon_0$ we have $c(\varepsilon) > 0$).

It follows from (15.4) and (15.7) that in the first case

$$\frac{\ln N_0^\Lambda(x_1, \dots, x_\ell; \varepsilon)}{\ell} \xrightarrow[\ell \rightarrow \infty]{P} 0 \quad (15.8)$$

for all $\varepsilon > 0$ and it follows from (15.3) and (15.7) that in the second case

$$\lim_{\ell \rightarrow \infty} P \left\{ \frac{\ln N_0^\Lambda(x_1, \dots, x_\ell; \varepsilon)}{\ell} > c(\varepsilon_0) - \delta \right\} = 1 \quad (15.9)$$

for all $\varepsilon \leq \varepsilon_0$, $\delta > 0$.

We will show that (15.8) implies uniform convergence on the means to their expectations, while under condition (15.9) such a convergence is not valid. Thus the following theorem is valid.

Theorem 15.1. *The equality*

$$\lim_{\ell \rightarrow \infty} \frac{H^\Lambda(\varepsilon, \ell)}{\ell} = 0, \quad \forall \varepsilon > 0$$

is a necessary and sufficient condition for the uniform convergence of means to their expectations for a bounded family of functions $F(x, \alpha)$ $\alpha \in \Lambda$.

This chapter is devoted to the proof of this theorem.

We now prove (as in Chapter 14) that the limit

$$\lim_{\ell \rightarrow \infty} \frac{H^\Lambda(\varepsilon, \ell)}{\ell} = c_0$$

exists and the convergence (15.8) is valid.

15.1.1 Proof of the Existence of the Limit

As

$$0 \leq \frac{H^\Lambda(\varepsilon, \ell)}{\ell} \leq 1,$$

for any $\varepsilon_0 > 0$ there is a lower bound

$$\liminf_{\ell \rightarrow \infty} \frac{H^\Lambda(\varepsilon_0, \ell)}{\ell} = c_0.$$

Therefore for any $\delta > 0$, such an ℓ_0 can be found that

$$\frac{H^\Lambda(\varepsilon_0, \ell_0)}{\ell_0} \leq c_0 + \delta.$$

Now take arbitrary $\ell > \ell_0$. Let

$$\ell = n\ell_0 + m,$$

where $n = [\ell/\ell_0]$. Then by virtue of (15.6) we obtain

$$\frac{H^\Lambda(\varepsilon_0, \ell)}{\ell} = \frac{H^\Lambda(\varepsilon_0, n\ell_0 + m)}{n\ell_0 + m} < \frac{nH^\Lambda(\varepsilon_0, \ell_0) + m}{n\ell_0} < \frac{H^\Lambda(\varepsilon_0, \ell_0)}{\ell_0} + \frac{1}{n}.$$

Strengthen the latter inequality

$$\frac{H^\Lambda(\varepsilon_0, \ell)}{\ell} < \frac{H^\Lambda(\varepsilon_0, \ell_0)}{\ell_0} + \frac{1}{n} < c_0 + \delta + \frac{1}{n}.$$

Since $n \rightarrow \infty$ when $\ell \rightarrow \infty$, we have

$$\limsup_{\ell \rightarrow \infty} \frac{H^\Lambda(\varepsilon_0, \ell)}{\ell} \leq c_0 + \delta.$$

Because $\delta > 0$ is arbitrary, the upper bound coincides with the lower one.

15.1.2 Proof of the Convergence of the Sequence

We prove that when ℓ increases, the sequence of random values

$$r^\ell = \frac{\ln N^\Lambda(x_1, \dots, x_\ell; \varepsilon_0)}{\ell}$$

converges in probability to the limit c_0 . For this it is sufficient to show that for any $\delta > 0$ we have

$$P_\delta^+(r^\ell) = P\{r^\ell > c_0 + \delta\} \xrightarrow[\ell \rightarrow \infty]{} 0$$

and for any $\mu > 0$ we obtain

$$P_\delta^-(r^\ell) = P\{r^\ell < c_0 - \mu\} \xrightarrow[\ell \rightarrow \infty]{} 0.$$

Consider a random sequence

$$g_n^{\ell_0} = \frac{1}{n} \sum_{i=1}^n r_i^{\ell_0}$$

of independent random values $r_i^{\ell_0}$. Evidently

$$E r_i^{\ell_0} = E g_n^{\ell_0} = \frac{H^\Lambda(\varepsilon_0, \ell_0)}{\ell_0}$$

Because $0 < r_i^{\ell_0} \leq 1$, we have

$$\begin{aligned} E(r_i^{\ell_0} - Er_i^{\ell_0})^2 &= D_2 \leq 1, \\ E(r_i^{\ell_0} - Er_i^{\ell_0})^4 &= D_4 \leq 1 \end{aligned}$$

Therefore

$$E(g_n^{\ell_0} - Eg_n^{\ell_0})^4 = \frac{D_4}{n^3} + 3\frac{n+1}{n^3}D_2 < \frac{4}{n^2}.$$

Write the Chebyshev's inequality for the fourth moment:

$$P \left\{ \left| g_n^{\ell_0} - \frac{H^\Lambda(\varepsilon_0, \ell_0)}{\ell_0} \right| > \varepsilon \right\} < \frac{4}{n^2 \varepsilon^4}.$$

Consider a random variable g_n^ℓ , where $\ell = n\ell_0 + m$. By virtue of (15.5),

$$r^\ell = r^{n\ell_0+m} \leq g_n^{\ell_0} + \frac{1}{n}.$$

Now let $\varepsilon = \delta/3$, and let ℓ_0 and $\ell = n\ell_0 + m$ be so large that

$$\frac{H^\Lambda(\varepsilon_0, \ell_0)}{\ell_0} - c_0 \leq \frac{\delta}{3}, \quad \frac{1}{n} \leq \frac{\delta}{3}.$$

Then

$$P_\delta^+(r^\ell) = P\{r^\ell - c_0 > \delta\} \leq P \left\{ \left| g_n^{\ell_0} - c_0 - \frac{2}{3}\delta \right| > \frac{\delta}{3} \right\} < \frac{244}{\delta^4 n^2}.$$

Because $n \rightarrow \infty$ when $\ell \rightarrow \infty$ we obtain

$$\lim_{\ell \rightarrow \infty} P_\delta^+(r^\ell) = 0.$$

To bound the value $P_\mu^-(r^\ell)$ we consider the equality

$$\int_0^{\frac{H^\Lambda(\varepsilon_0, \ell)}{\ell}} \left(\frac{H^\Lambda(\varepsilon_0, \ell)}{\ell} - r^\ell \right) dP(r^\ell) = \int_{\frac{H^\Lambda(\varepsilon_0, \ell)}{\ell}}^1 \left(r^\ell - \frac{H^\Lambda(\varepsilon_0, \ell)}{\ell} \right) dP(r^\ell)$$

Mark its left-hand side with R_1 , mark its right-hand side with R_2 , and bound R_1 and R_2 for ℓ such that

$$\frac{H^\Lambda(\varepsilon_0, \ell)}{\ell} - c_0 < \frac{\mu}{2}.$$

The lower bound of R_1 is

$$R_1 = \frac{\frac{H^\Lambda(\varepsilon_0, \ell)}{\ell} - r^\ell}{\ell} dP(r^\ell) \geq \frac{\mu}{2} \int_0^{c_0 - \mu} dP(r^\ell) = \frac{\mu}{2} R_\mu^-(r^\ell)$$

and the upper bound of R_2 is

$$\begin{aligned} R_2 &= \int_{\frac{H^\Lambda(\varepsilon_0, \ell)}{\ell}}^{c_0 + \delta} \left(r^\ell - \frac{H^\Lambda(\varepsilon_0, \ell)}{\ell} \right) dP(r^\ell) + \int_{c_0 + \delta}^1 \left(r^\ell - \frac{H^\Lambda(\varepsilon_0, \ell)}{\ell} \right) dP(r^\ell) \\ &\leq \left| c_0 + \delta - \frac{H^\Lambda(\varepsilon_0, \ell)}{\ell} \right| + P_\delta^+(r^\ell). \end{aligned}$$

Combining these bounds we obtain

$$\frac{\mu}{2} P_\mu^-(r^\ell) \leq \left| c_0 + \delta - \frac{H^\Lambda(\varepsilon_0, \ell)}{\ell} \right| + P_\delta^+(r^\ell).$$

Since

$$\begin{aligned} \frac{H^\Lambda(\varepsilon_0, \ell)}{\ell} &\xrightarrow[\ell \rightarrow \infty]{} c_0, \\ P_\delta^+(r^\ell) &\xrightarrow[\ell \rightarrow \infty]{} 0, \end{aligned}$$

we obtain

$$\lim_{\ell \rightarrow \infty} P_\mu^-(r^\ell) \leq \frac{2\delta}{\mu}.$$

Because δ and μ are arbitrary, we conclude that

$$\lim_{\ell \rightarrow \infty} P_\mu^-(r^\ell) \xrightarrow[\ell \rightarrow \infty]{} 0$$

15.2 THE QUASICUBE

We shall define by induction an n -dimensional quasicube with an edge a .

Definition. A set \mathcal{Q} in the space R^1 is called a one-dimensional quasicube with an edge a if \mathcal{Q} is a segment $[c, c+a]$.

A set \mathcal{Q} in the space R^n is called an n -dimensional quasicube with an edge a if there exists a coordinate subspace R^{n-1} (for simplicity it will be assumed below that this subspace is formed by the first $n-1$ coordinates) such that a projection $\bar{\mathcal{Q}}$ of the set \mathcal{Q} on this subspace is an $(n-1)$ -dimensional quasicube

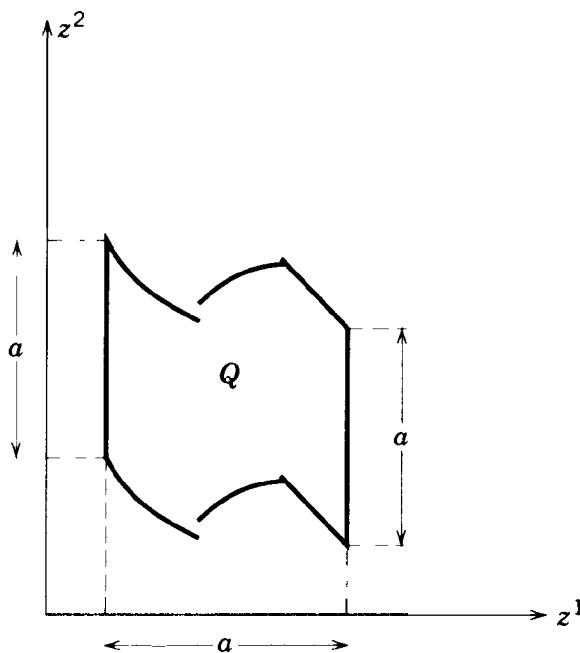


FIGURE 15.1.

with an edge a and for each point $z_* = (z_*^1, \dots, z_*^{n-1})$ of the quasicube \bar{Q} the set of numerical values z^n such that $(z_*^1, \dots, z_*^{n-1}, z^n) \in Q$ forms a segment $[c\zeta + a]$, where c in general does not depend on z_* . (Fig. 15.1).

The space R^{n-1} is called an $(n-1)$ -dimensional *canonical* space. In turn, an $(n-2)$ -dimensional canonical space R^{n-2} can be constructed for this space, and so on.

The totality of subspaces R^1, \dots, R^n is called a *canonical structure*.

The following lemma is valid. This lemma is an analog (for the value of the set) of the Lemma 4.1 proved in Chapter 4, Section 4.10.

Lemma 15.1. *Let a convex set A belong to an ℓ -dimensional cube whose coordinates satisfy*

$$0 \leq z^i \leq 1, \quad i = 1, \dots, \ell.$$

Let $V(A)$ be the ℓ -dimensional volume of the set A .

If for some

$$1 \leq n \leq \ell, \quad 0 \leq a \leq 1, \quad \ell > 1$$

the condition

$$V(A) > C_\ell^n a^{\ell-n} \tag{15.10}$$

is fulfilled, one can then find a coordinate n -dimensional subspace such that the projection of the set A on this subspace contains a quasicube with an edge a .

Proof. We shall prove the lemma using an induction method.

1. For $n = \ell$ the condition (15.10) is

$$V(A) > C_\ell^n = 1. \quad (15.11)$$

On the other hand,

$$V(A) \leq 1. \quad (15.12)$$

Therefore the condition (15.10) is never fulfilled and the assertion of the lemma is trivially valid.

2. For $n = 1$ and any ℓ we shall prove the lemma by contradiction. Let there exist no one-dimensional coordinate space such that the projection of the set A on this space contains the segment $[c, c + a]$. The projection of a bounded convex set on the one-dimensional axis is either an open interval, a segment, or a semiclosed interval. Consequently, by assumption the length of this interval does not exceed a . However, then the set A itself is contained in an (ordinary) cube with an edge a . This implies that

$$V(A) < a'.$$

Taking into account that $a \leq 1$, we obtain

$$V(A) < a^\ell < \ell a^{\ell-1},$$

which contradicts condition (15.10) of the lemma.

3. Consider now the general inductive step. Let the lemma be valid for all $n < n_0$ for all ℓ , as well as for $n = n_0 + 1$ for all ℓ such that $n \leq \ell \leq \ell_0$. We shall show that it is valid for $n = n_0 + 1$, $\ell = \ell_0 + 1$.

Consider a coordinate subspace R^{ℓ_0} of dimension ℓ_0 consisting of vectors

$$z = (z^1, \dots, z^{\ell_0}).$$

Let A' be a projection of A on this subspace. (Clearly A' is convex.) If

$$V(A') > C_{\ell_0}^n a^{\ell_0-n}, \quad (15.13)$$

then by the induction assumption there exists a subspace of dimension n such that the projection of the set A' on this subspace contains a quasicube with an edge a . The lemma is thus proved in the case (15.13).

Let

$$V(A') \leq C_{\ell_0}^n a^{\ell_0-n}. \quad (15.14)$$

Consider two functions

$$\phi_1(z^1, \dots, z^{\ell_0}) = \sup\{z : (z^1, \dots, z^{\ell_0}, z) \in A\},$$

$$\phi_2(z^1, \dots, z^{\ell_0}) = \inf_z\{z : (z^1, \dots, z^{\ell_0}, z) \in A\}.$$

These functions are convex upward and downward, respectively. Therefore the function

$$\phi_3(z^1, \dots, z^{\ell_0}) = \phi_1(z^1, \dots, z^{\ell_0}) - \phi_2(z^1, \dots, z^{\ell_0})$$

is convex upward.

Consider the set

$$A^{II} = \{(z^1, \dots, z^{\ell_0}) : \phi_3(z^1, \dots, z^{\ell_0}) > a\} \quad (15.15)$$

This set is convex and is located in R^{ℓ_0} .

For the set A'' , one of two inequalities is fulfilled: Either

$$V(A^{II}) > C_{\ell_0}^{n-1} a^{\ell_0-n+1} \quad (15.16)$$

or

$$V(A^{II}) \leq C_{\ell_0}^{n-1} a^{\ell_0-n+1}. \quad (15.17)$$

Assume that (15.16) is fulfilled. Then by the induction assumption there exists a coordinate space R^{n-1} of the space R^ℓ such that projection A''' of the set A'' on it contains an $(n-1)$ -dimensional quasicube Ω_{n-1} with an edge a .

We now consider the n -dimensional coordinate subspace R^n formed by R^{n-1} and the coordinate z^{ℓ_0} . Let, A^{IV} be the projection of the set A on the subspace R^n . For a given point

$$(z_*^1, \dots, z_*^{n-1}) \in A'''$$

we consider the set $d = d(z_*^1, \dots, z_*^{n-1})$ of values of z such that

$$(z_*^1, \dots, z_*^{n-1}, z) \in A^{IV}.$$

It is easy to see that the set d contains an interval with endpoints

$$r_1(z^1, \dots, z^{n-1}) = \sup_{z \in A^{II}} \phi_1(z^1, \dots, z^{\ell_0}),$$

$$r_2(z^1, \dots, z^{n-1}) = \inf_{z \in A^{II}} \phi_2(z^1, \dots, z^{\ell_0}),$$

where \sup^* and \inf^* are taken over the points $z \in A^{II}$ which are projected onto a given point $(z_*^1, \dots, z_*^{n-1})$. Clearly, in view of (15.15) we have

$$r_1 - r_2 > a.$$

We now assign to each point $(z^1, \dots, z^{n-1}) \in A^{III}$ a segment $c(z^1, \dots, z^{n-1})$ of length a on the axis z^{ℓ_0+1} :

$$\left[\sigma - \frac{a}{2}, \sigma + \frac{a}{2} \right],$$

where

$$\sigma = \frac{r_1(z^1, \dots, z^{n-1}) + r_2(z^1, \dots, z^{n-1})}{2}$$

Clearly,

$$c(z^1, \dots, z^{n-1}) \subset d(z^1, \dots, z^{n-1}).$$

Consider now the set $Q \in R^n$ consisting of points $(z^1, \dots, z^{n-1}, z^{\ell_0+1})$ such that

$$(z^1, \dots, z^{n-1}) \in \Omega_{n-1}, \quad (15.18)$$

$$z^{\ell_0+1} \in c(z^1, \dots, z^{n-1}). \quad (15.19)$$

This set is the required quasicube Ω_n . Indeed, in view of (15.18) and (15.19) the set Q satisfies the definition of an n -dimensional quasicube with an edge a . At the same time we have $Q \in A^{IV}$ by construction.

To prove the lemma, we need to consider the case when the inequality (15.17) is fulfilled, that is,

$$V(A^{II}) \leq C_{\ell_0}^{n-1} a^{\ell_0-n+1}.$$

Then

$$\begin{aligned} V(A) &= \int_{A'} \phi_3(z^1, \dots, z^{\ell_0}) dz^1 dz^2 \dots dz^{\ell_0} \\ &= \int_{A' - A''} \phi_3(z^1, \dots, z^{\ell_0}) dz^1 dz^2 \dots dz^{\ell_0} \\ &\quad + \int_{A''} \phi_3(z^1, \dots, z^{\ell_0}) dz^1 dz^2 \dots dz^{\ell_0} \\ &\leq a V(A') + V(A''), \end{aligned}$$

and in view of (15.14) and (15.17) we obtain

$$V(A) \leq C_{\ell_0}^n a^{\ell_0-n+1} + C_{\ell_0}^{n-1} a^{\ell_0-n+1} = C_{\ell_0+1}^n a^{\ell_0-n+1},$$

which contradicts the lemma's condition.

15.3 ε -EXTENSION OF A SET

Let A be a convex bounded set in R^n . We assign to each point $z \in A$ an open cube $\Omega(z)$ with the center at z and the edge ε oriented along the coordinate axes.

Consider the set

$$A_\varepsilon = \bigcup_{z \in A} \Omega(z),$$

along with the set A , which we shall call an ε extension of the set A . The set A_ε is the set of points $y = (y^1, \dots, y^\ell)$, for each of which there exists a point $z \in A$ such that

$$\rho(z, y) < \frac{\varepsilon}{2}.$$

It is easy to show that an ε extension A_ε of the convex set A is convex.

Now choose a minimal proper ε net on the set A . Let the minimal number of elements of a proper ε net of the set A be $N_0(\varepsilon, A)$. Denote by $V(A_\varepsilon)$ the volume of the set A_ε .

Lemma 15.2. The inequality

$$N_0(1.5\varepsilon, A)\varepsilon^\ell \leq V(A_\varepsilon) \quad (15.20)$$

is valid.

Proof. Let T be a proper $\varepsilon/2$ net of the set A . Select a subset \hat{T} of the set T according to the following rules:

1. The first point \hat{z}_1 of the set \hat{T} is an arbitrary point of T .
2. Let m distinct points $\hat{z}_1, \dots, \hat{z}_m$ be chosen. An arbitrary point of $z \in T$ such that

$$\min_{1 \leq i \leq m} \rho(\hat{z}_i, z) \geq \varepsilon$$

is selected as an $(m+1)$ th point of \hat{T} .

3. If there is no such point or if T has been exhausted, then the construction is completed.

Let the set \hat{T} , constructed in the manner described above, be a 1.5ε net in A . Indeed, for any $z \in A$, there exists $t \in T$ such that $\rho(z, t) < \varepsilon/2$. For such a t there exists $\hat{z} \in \hat{T}$ such that $\rho(\hat{z}, t) < \varepsilon$. Consequently, $\rho(z, \hat{z}) < 1.5\varepsilon$ and the number of elements in T is at least $N_0(1.5\varepsilon, A)$.

Furthermore, the union of open cubes with edge ε and centers at the points of \hat{T} is included in A_ε . At the same time, cubes with centers at

different points do not intersect. (Otherwise, there would exist $\hat{z} \in \Omega(z_1)$ and $\hat{z} \in \Omega(z_2)$, $z_1, z_2 \in \hat{T}$, and hence $\rho(z_1, \hat{z}) < \varepsilon/2$ and $\rho(z_2, \hat{z}) < \varepsilon/2$, from which $\rho(z_1, z_2) < \varepsilon$ and $z_1 = z_2$.) Consequently,

$$V(A_\varepsilon) \geq N_0(1.5\varepsilon, A)\varepsilon^\ell.$$

The lemma is proved.

Lemma 153. Let a convex set A belong to the unit cube in R^ℓ , and let A_ε be its ε -extension ($0 < \varepsilon \leq 1$); and for some

$$\gamma > \ln(1 + \varepsilon)$$

let the inequality

$$N_0(1.5\varepsilon, A) > e^{\gamma\ell}$$

be fulfilled. Then there exist $t(\varepsilon, \gamma)$ and $a(\varepsilon, \gamma)$ such that—provided that $n = [t_0\ell] > 0$ —one can find a coordinate subspace of dimension $n = [t_0\ell]$ such that a projection of A_ε on this space contains an n -dimensional quasicube with an edge a .

Proof. In view of Lemmas 15.1 and 15.2 and the condition (15.20), which is valid for this lemma, in order that there exist an n -dimensional coordinate subspace such that the projection of A_ε on this subspace would contain an n -dimensional quasicube with an edge a , it is sufficient that

$$C_\ell^n b^{\ell-n} < e^{\gamma\ell} \varepsilon^\ell (1 + \varepsilon)^{-\ell},$$

where $b = a/(1 + \varepsilon)$.

In turn it follows from Stirling's formula that for this purpose it is sufficient that

$$b^{\ell-n} \left(\frac{e\ell}{n} \right)^n < e^{\gamma_1\ell} \varepsilon^\ell,$$

where

$$\gamma_1 = \gamma \ln(1 + \varepsilon).$$

Setting $t = n/\ell$ and taking $0 < t < \frac{1}{3}$, we obtain

$$-\frac{t(\ln t - 1)}{1-t} + \ln b < \frac{\ln \varepsilon + \gamma_1}{1-t},$$

using an equivalent transformation.

Under the stipulated restrictions this equality will be fulfilled if the inequality

$$-\frac{3}{2}t(\ln t - 1) + \ln b < (1 + 2t)\ln \varepsilon + \frac{2}{3}\gamma_1 \quad (15.21)$$

is satisfied. Now choose $t_0(\varepsilon, \gamma)$ such that conditions

$$\begin{aligned} 0 &< t_0(\varepsilon, \gamma) \leq \frac{1}{3}, \\ -\frac{3}{2}t_0(\ln t_0 - 1) &< \frac{\gamma_1}{6}, \\ -2t_0 \ln \varepsilon &< \frac{\gamma_1}{6} \end{aligned}$$

will be satisfied. This can always be achieved, since by assumption in this case the inequality (15.21) will be fulfilled for

$$\ln b = \ln \varepsilon + \frac{\gamma_1}{3},$$

$$a = (1 + \varepsilon)\varepsilon \exp \left\{ \frac{\gamma - \ln(1 - \varepsilon)}{3} \right\}. \quad (15.22)$$

The lemma is thus proved.

15.4 AN AUXILIARY LEMMA

Now consider a class of functions $\Phi = \{F(x, a) : a \in A\}$, which is defined on X . We assume the class to be convex in the sense that if

$$F(x, \alpha_1), \dots, F(x, \alpha_r) \subset \Phi, \quad (15.23)$$

then

$$\sum_{i=1}^r \tau_i F(x, \alpha_i) \subset \Phi, \quad \sum_{i=1}^r \tau_i = 1, \quad \tau_i \geq 0.$$

Now define two sequences: the sequence

$$x_1, \dots, x_\ell, \quad x_i \in X$$

and a random independent numerical sequence

$$(15.24)$$

possessing the property

$$y_i = \begin{cases} 1 & \text{with probability 0.5,} \\ -1 & \text{with probability 0.5.} \end{cases}$$

Using these sequences, we define the quantity

$$\mathcal{Q}(\Phi) = E_y \sup_{F(x,\alpha) \in \Phi} \frac{1}{\ell} \left| \sum_{i=1}^{\ell} F(x_i, \alpha) y_i \right|.$$

(The expectation is taken over the random sequences (15.24).)

In Section 15.1 we denoted by A the set of ℓ -dimensional vectors z with coordinates

$$z^i = F(x_i, \alpha), \quad i = 1, \dots, \ell,$$

for all possible $\alpha \in A$. Clearly A belongs to the unit ℓ -dimensional cube in R^ℓ and is convex.

We rewrite the function $\mathcal{Q}(\Phi)$ in the form

$$\mathcal{Q}(\Phi) = E_y \sup_{F(x,\alpha) \in \Phi} \frac{1}{\ell} \left| \sum_{i=1}^{\ell} z^i y_i \right|.$$

The following lemma is valid.

Lemma 15.4. If for $\varepsilon > 0$ the inequality

$$N_0(1.5\varepsilon, A) > e^{\gamma\ell}, \quad \gamma > \ln(1 + \varepsilon)$$

is fulfilled for the set A , then the inequality

$$\mathcal{Q}(\Phi) \geq \varepsilon \left(\exp \left\{ \frac{\gamma - \ln(1 + \varepsilon)}{3} \right\} - 1 \right) \left(\frac{t}{2} - \frac{1}{2\ell} \right)$$

is valid, where $t > 0$ does not depend on ℓ .

Proof. As was shown in the previous section, if the conditions of the lemma are fulfilled, there exist $t(\varepsilon, y)$ and $a(\varepsilon, y)$ such that there exists a coordinate subspace of dimension $n = [t\ell]$ with the property that a projection of the set A , on this subspace contains an n -dimensional quasicube with edge a . We have assumed here, without loss of generality, that this subspace forms the first n coordinates and that the corresponding n -dimensional subspace forms a canonical subspace of this quasicube.

We define the vertices of the quasicube using the following iterative rule:

1. The vertices of the one-dimensional cube are the endpoints of the segment c and $c + a$.

2. To define vertices of an n-dimensional quasicube in an n-dimensional canonical space, we proceed as follows. Let the vertices of an $(n - 1)$ -dimensional quasicube be determined. Assign the segment

$$\left[\phi^{n-1}(\hat{z}_k^1, \dots, \hat{z}_k^{n-1}) - \frac{a}{2}, \phi^{n-1}(\hat{z}_k^1, \dots, \hat{z}_k^{n-1}) + \frac{a}{2} \right]$$

to each such vertex $\hat{z}_k^1, \dots, \hat{z}_k^{n-1}$ (k is number of the vertex), where

$$\begin{aligned}\phi^{n-1}(\hat{z}_k^1, \dots, \hat{z}_k^{n-1}) &= \frac{1}{2} (\phi_1(\hat{z}_k^1, \dots, \hat{z}_k^{n-1}) + \phi_2(\hat{z}_k^1, \dots, \hat{z}_k^{n-1})), \\ \phi_1(\hat{z}_k^1, \dots, \hat{z}_k^{n-1}) &= \max_{\hat{z}^n} \left\{ \hat{z}^n : (\hat{z}^1, \dots, \hat{z}^{n-1}, \hat{z}^n) \in \Omega_n \right\}, \\ \phi_2(\hat{z}_k^1, \dots, \hat{z}_k^{n-1}) &= \min_{\hat{z}^n} \left\{ \hat{z}^n : (\hat{z}^1, \dots, \hat{z}^{n-1}, \hat{z}^n) \in \Omega_n \right\},\end{aligned}$$

and Ω_n is an n-dimensional quasicube.

This segment is formed by the intersection of the line $(\hat{z}_k^1, \dots, \hat{z}_k^{n-1}, z^n)$ and the quasicube. The endpoints of the segment form the vertices of the quasicube.

Thus if

$$(\hat{z}_k^1, \dots, \hat{z}_k^{n-1}) \in R^{n-1}$$

is the k th vertex of an $(n - 1)$ -dimensional quasicube, then

$$\begin{aligned}&\left(\hat{z}_k^1, \dots, \hat{z}_k^{n-1}, \phi^{n-1}(\hat{z}_k^1, \dots, \hat{z}_k^{n-1}) + \frac{a}{2} \right), \\ &\left(\hat{z}_k^1, \dots, \hat{z}_k^{n-1}, \phi^{n-1}(\hat{z}_k^1, \dots, \hat{z}_k^{n-1}) - \frac{a}{2} \right)\end{aligned}$$

are correspondingly the $(2k - 1)$ th and $2k$ th vertices of the n-dimensional quasicube.

Now we assign to an arbitrary sequence

$$y_1, \dots, y_n \quad (y_i = \{1, -1\})$$

a vertex \hat{z}_* of a quasicube defined as follows:

$$\begin{aligned}\hat{z}_*^1 &= \left(c + \frac{a}{2} \right) + \frac{a}{2} y_1, \\ \hat{z}_*^j &= \phi^{j-1}(\hat{z}_*^1, \dots, \hat{z}_*^{j-1}) + \frac{a}{2} y_j, \quad j = 2, \dots, n.\end{aligned}$$

In turn, to each vertex \hat{z}_* of a quasicube in R^n we assign a point

$$z_* = (z_*^1, \dots, z_*^\ell) \in A$$

such that the distance between the projection (z_*^1, \dots, z_*^n) of this point in R^n and the vertex \hat{z}_* is at most $\varepsilon/2$, that is,

$$|z_*^j - \hat{z}_*^j| < \frac{\varepsilon}{2}, \quad j = 1, \dots, n.$$

This is possible because $z_* \in \text{Pr } A$, on R^n .

Thus we introduce two functions

$$\begin{aligned}\hat{z}_* &= \hat{z}_*(y_1, \dots, y_n), \\ z_* &= z_*(\hat{z}_*^1, \dots, \hat{z}_*^n).\end{aligned}$$

We shall denote the difference $z_*^j - \hat{z}_*^j$ by 8_j ($j = 1, \dots, n$) ($|8_j| \leq \varepsilon/2$) and bound the quantity

$$\begin{aligned}Q(\Phi) &= E \sup_{z \in A} \frac{1}{\ell} \left| \sum_{i=1}^{\ell} z^i y_i \right| \\ &\geq \frac{1}{\ell} E \sum_{i=1}^{\ell} z_*^i y_i \\ &= \frac{1}{\ell} \sum_{i=1}^n E y_i (\hat{z}_*^i + \delta_i) + \frac{1}{\ell} \sum_{i=n+1}^{\ell} E y_i z_*^i.\end{aligned}$$

Observe that the second summand in the sum is zero, since every term of the sum is a product of two independent random variables y_i and z_*^i , $i > n$, one of which (y_i) has zero mean.

We shall bound the first summand. For this purpose consider the first term in the first summand:

$$\begin{aligned}\frac{1}{\ell} E \left[y_1 \left(c + \frac{a}{2} + \frac{a}{2} y_1 + \delta_1 \right) \right] &= \frac{1}{\ell} \left[\frac{a}{2} + E y_1 \delta_1 \right] \\ &\geq \frac{1}{2\ell} (a - \varepsilon).\end{aligned}$$

To bound the k th term

$$I_k = \frac{1}{\ell} E \left[y_k \left(\phi^{k-1}(\hat{z}_*^1, \dots, \hat{z}_*^{k-1}) + \frac{a}{2} y_k + \delta_k \right) \right],$$

we observe that the vertex $(\hat{z}_*^1, \dots, \hat{z}_*^{k-1})$ was chosen in such a manner that it would not depend on y_k but only on y_1, \dots, y_{k-1} . Therefore

$$I_k = \frac{1}{\ell} \left[\frac{a}{2} + E y_k \delta_k \right] \geq \frac{1}{2\ell} (a - \varepsilon).$$

Thus we obtain

$$\mathcal{Q}(\Phi) > E \sup_{z_* \in A} \frac{1}{\ell} \sum_{i=1}^{\ell} z'_* y_i \geq \frac{n}{2\ell} (a - \varepsilon) > (a - \varepsilon) \left(\frac{t}{2} - \frac{1}{2\ell} \right).$$

Choosing the quantity a in accordance with (15.22), we arrive at

$$\mathcal{Q}(\Phi) > \varepsilon \left(\exp \left\{ \frac{\gamma - \ln(1 + \varepsilon)}{3} \right\} - 1 \right) \left(\frac{t}{2} - \frac{1}{2\ell} \right)$$

The lemma is thus proved.

15.5 NECESSARY AND SUFFICIENT CONDITIONS FOR UNIFORM CONVERGENCE. THE PROOF OF NECESSITY

Theorem 15.1. *For the uniform convergence of the means to their mathematical expectations over a uniformly bounded class of functions $F(x, \alpha), \alpha \in A$, it is necessary and sufficient that for any $\varepsilon > 0$ the equality*

$$\lim_{\ell \rightarrow \infty} \frac{H^A(\varepsilon, f?)}{\varepsilon} = 0 \quad (15.25)$$

be satisfied.

To prove the necessity we can assume, without loss of generality, that the class $F(x, \alpha), \alpha \in A$, is convex in the sense of (15.23), since from the uniform convergence of the means to their mathematical expectations for an arbitrary class, it follows that we obtain the same convergence for its convex closure, and the condition (15.25) for a convex closure implies the same for the initial class of functions.

Proof of Necessity. Assume the contrary. For some $\varepsilon_0 > 0$ let the equality

$$\lim_{\ell \rightarrow \infty} \frac{H^A(\varepsilon_0, \ell)}{\varepsilon_0} = c(\varepsilon_0) > 0 \quad (15.26)$$

be fulfilled, and at the same time let uniform convergence hold; that is, for all ε let the relationship

$$\lim_{\ell \rightarrow \infty} P \left\{ \sup_{\alpha \in A} \left| EF(x, \alpha) - \frac{1}{\ell} \sum_{i=1}^{\ell} F(x_i, \alpha) \right| > \varepsilon \right\} = 0 \quad (15.27)$$

be satisfied. This will lead to a contradiction.

Since functions, $F(x, \alpha)$, $\alpha \in \Lambda$, are uniformly bounded by 1, it follows from (15.27) that

$$\lim_{\ell \rightarrow \infty} E \left\{ \sup_{\alpha \in \Lambda} \left| EF(x, \alpha) - \frac{1}{\ell} \sum_{i=1}^{\ell} F(x_i, \alpha) \right| \right\} = 0.$$

This implies that if $\ell_1 \rightarrow \infty$ and $\ell - \ell_1 \rightarrow \infty$, then the equality

$$\lim_{\ell_1, \ell \rightarrow \infty} E \left\{ \sup_{\alpha \in \Lambda} \left| \frac{1}{\ell_1} \sum_{i=1}^{\ell_1} F(x_i, \alpha) - \frac{1}{\ell - \ell_1} \sum_{i=\ell_1+1}^{\ell} F(x_i, \alpha) \right| \right\} = 0 \quad (15.28)$$

is fulfilled.

Consider the expression

$$I(x_1, \dots, x_\ell) = \sum_{n=0}^{\ell} \sup_{\alpha \in \Lambda} \left[\frac{C_\ell^n}{2^\ell} \frac{1}{\ell} \left| \sum_{i=1}^n F(x_i, \alpha) - \sum_{i=n+1}^{\ell} F(x_i, \alpha) \right| \right]$$

We subdivide the summation with respect to n into two "regions"

$$\begin{aligned} I: \quad & \left| n - \frac{\ell}{2} \right| < \ell^{2/3}, \\ II: \quad & \left| n - \frac{\ell}{2} \right| \geq \ell^{2/3}. \end{aligned}$$

Then taking into account that

$$\frac{1}{\ell} \left| \sum_{i=1}^n F(x_i, \alpha) - \sum_{i=n+1}^{\ell} F(x_i, \alpha) \right| \leq 1,$$

we obtain

$$\begin{aligned} I(x_1, \dots, x_\ell) & \leq \sum_{n \in II} \frac{C_\ell^n}{2^\ell} + \sum_{n \in I} \frac{C_\ell^n}{2^\ell} \sup_{\alpha \in \Lambda} \left| \frac{n}{\ell} \cdot \left(\frac{1}{n} \sum_{i=1}^n F(x_i, \alpha) \right) \right. \\ & \quad \left. - \frac{\ell - n}{\ell} \left(\frac{1}{\ell - n} \sum_{i=n+1}^{\ell} F(x_i, \alpha) \right) \right|. \end{aligned}$$

Note that in region I ($1/2 - 1/\ell^{2/3} < n/\ell < 1/2 + 1/\ell^{2/3}$) we have

$$\sum_{n \in I} \frac{C_\ell^n}{2^\ell} \xrightarrow{\ell \rightarrow \infty} 1,$$

while in region 11 we obtain

$$\sum_{n \in II} \frac{C_\ell^n}{2^\ell} \xrightarrow{\ell \rightarrow \infty} 0. \quad (15.29)$$

Furthermore,

$$\begin{aligned} & \lim_{\ell \rightarrow \infty} EI(x_1, \dots, x_\ell) \\ & \leq \lim_{\ell \rightarrow \infty} \left(\sum_{n \in I} \frac{C_\ell^n}{2^\ell} + \frac{1}{2} \max_{n \in I} E \sup_{\alpha \in \Lambda} \left| \frac{1}{n} \sum_{i=1}^n F(x_i, \alpha) - \frac{1}{\ell-n} \sum_{i=n+1}^\ell F(x_i, \alpha) \right| \sum_{n \in I} \frac{C_\ell^n}{2^\ell} \right). \end{aligned}$$

It follows from (15.28) that

$$\max_{n \in I} E \sup_{\alpha \in \Lambda} \left| \frac{1}{n} \sum_{i=1}^n F(x_i, \alpha) - \frac{1}{\ell-n} \sum_{i=n+1}^\ell F(x_i, \alpha) \right| \xrightarrow{\ell \rightarrow \infty} 0.$$

Thus taking (15.29) into account we have

$$\lim_{\ell \rightarrow \infty} EI(x_1, \dots, x_\ell) = 0. \quad (15.30)$$

On the other hand,

$$EI(x_1, \dots, x_\ell) = E \frac{1}{\ell!} \sum_{k=1}^{\ell!} I(T_k \{x_1, \dots, x_\ell\}),$$

where T_k ($k = 1, \dots, \ell!$) are all the permutations of the sequence. We transform the right-hand side:

$$\begin{aligned} & E \frac{1}{\ell!} \sum_{k=1}^{\ell!} I(T_k \{x_1, \dots, x_\ell\}) \\ & = E \frac{1}{\ell!} \sum_{k=1}^{\ell!} \sum_{n=0}^{\ell} \sup_{\alpha \in \Lambda} \left[\frac{C_\ell^n}{2^\ell} \frac{1}{\ell} \left| \sum_{i=1}^n F(x_{j(i,k)}, \alpha) - \sum_{i=n+1}^\ell F(x_{j(i,k)}, \alpha) \right| \right] \\ & = E \sum_{n=0}^{\ell} \frac{1}{C_\ell^n} \sum_{y_1, \dots, y_\ell} \sup_{\alpha \in \Lambda} \frac{C_\ell^n}{2^\ell} \frac{1}{\ell} \left| \sum_{i=1}^{\ell} y_i F(x_i, \alpha) \right|. \end{aligned}$$

(Here $j(i, k)$ is the index obtained when the permutation T_k acts on i .) In the last expression the summation is carried out over all the sequences

$$y_1, \dots, y_\ell, \quad y_i = \begin{cases} 1, \\ -1, \end{cases}$$

which have n positive values.

Furthermore, we obtain

$$EI(x_1, \dots, x_\ell) = E \frac{1}{2^\ell} \left\{ \sum_{y_1, \dots, y_\ell} \sup_{\alpha \in \Lambda} \frac{1}{\ell} \left| \sum_{i=1}^{\ell} y_i F(x_i, \alpha) \right| \right\}. \quad (15.31)$$

In (15.31) the summation is carried over all sequences

$$y_1, \dots, y_\ell.$$

Choose for $\varepsilon_0 > 0$ a number such that

$$\lim_{\ell \rightarrow \infty} \frac{H^\Lambda(\varepsilon_0, \ell)}{\ell} = c(\varepsilon_0) > 0.$$

Since $c(\varepsilon)$ is nondecreasing as ε decreases, one can choose ε so that the relations

$$\begin{aligned} 0 < 1.5\varepsilon \leq \varepsilon_0, \\ \ln(1 + \varepsilon) < \frac{c(\varepsilon_0) \ln 2}{2}, \\ c(1.5\varepsilon) \geq c(\varepsilon_0) \end{aligned}$$

are fulfilled. Then in view of (15.9) the probability of fulfillment of the inequality

$$N_0^\Lambda(x_1, \dots, x_\ell; 1.5\varepsilon) > \exp \left\{ \frac{c(\varepsilon_0) \ln 2}{2} \right\} \ell \quad (15.32)$$

tends to 1. According to Lemma 15.4, when (15.32) is satisfied, the expression in the curly brackets in (15.31) exceeds the quantity

$$\varepsilon \left(\exp \left\{ \frac{\gamma}{3} \right\} - 1 \right) \left(\frac{t}{2} - \frac{1}{2\ell} \right),$$

where

$$\gamma = \frac{c(\varepsilon_0) \ln 2}{2} - \ln(1 + \varepsilon)$$

and $t = t(\varepsilon, y)$ does not depend on ℓ . Hence we conclude that

$$\lim_{\ell \rightarrow \infty} I(x_1, \dots, x_\ell) > \lim_{\ell \rightarrow \infty} \varepsilon \left(\exp \left\{ \frac{\gamma}{3} \right\} - 1 \right) \left(\frac{t}{2} - \frac{1}{2\ell} \right) > 0.$$

This inequality contradicts the statement (15.30). The contradiction obtained proves the first part of the theorem.

15.6 NECESSARY AND SUFFICIENT CONDITIONS FOR UNIFORM CONVERGENCE. THE PROOF OF SUFFICIENCY

The following lemma is valid.

Lemma 15.5. *For $\ell > \frac{2}{\varepsilon^2}$ the inequality*

$$\begin{aligned} P \left\{ \sup_{\alpha \in \Lambda} \left| \frac{1}{\ell} \sum_{i=1}^{\ell} F(x_i, \alpha) - EF(x, \alpha) \right| > \varepsilon \right\} \\ \leq 2P \left\{ \sup_{\alpha \in \Lambda} \left| \frac{1}{\ell} \sum_{i=1}^{\ell} F(x_i, \alpha) - \frac{1}{\ell} \sum_{i=\ell+1}^{2\ell} F(x_i, \alpha) \right| > \frac{\varepsilon}{2} \right\} \end{aligned}$$

holds true.

Therefore if for any $\varepsilon > 0$ the relation

$$P \left\{ \sup_{\alpha \in \Lambda} \left| \frac{1}{\ell} \sum_{i=1}^{\ell} F(x_i, \alpha) - \frac{1}{\ell} \sum_{i=\ell+1}^{2\ell} F(x_i, \alpha) \right| > \varepsilon \right\} \xrightarrow{\ell \rightarrow \infty} 0 \quad (15.33)$$

is valid, then for any $\varepsilon > 0$ the convergence

$$P \left\{ \sup_{\alpha \in \Lambda} \left| \frac{1}{\ell} \sum_{i=1}^{\ell} F(x_i, \alpha) - EF(x, \alpha) \right| > \varepsilon \right\} \xrightarrow{\ell \rightarrow \infty} 0$$

also holds true.

Proof. The proof of this lemma mainly repeats the proof of the basic lemma (Section 14.2). The only difference is that instead of Chernoff's inequality we use Hoeffding's inequality. We denote by R_ℓ the event

$$\left\{ \sup_{\alpha \in \Lambda} \left| \frac{1}{\ell} \sum_{i=1}^{\ell} F(x_i, \alpha) - EF(x, \alpha) \right| > \varepsilon_0 \right\}.$$

Then for sufficiently large ℓ the inequality

$$P\{R_\ell\} > \eta > 0$$

is fulfilled. We introduce the notation

$$\frac{1}{\ell} \left| \sum_{i=1}^{\ell} F(x_i, \alpha) - \sum_{i=\ell+1}^{2\ell} F(x_i, \alpha) \right| = S(x_1, \dots, x_{2\ell}; \alpha)$$

and consider the quantity

$$\begin{aligned} P_{2\ell} &= P \left\{ \sup_{\alpha \in \Lambda} S(x_1, \dots, x_{2\ell}; \alpha) > \frac{\varepsilon_0}{3} \right\} \\ &= \int_{x_1} \dots \int_{x_{2\ell}} \theta \left[\sup_{\alpha \in \Lambda} S(x_1, \dots, x_{2\ell}; \alpha) - \frac{\varepsilon_0}{3} \right] dP(x_1) \dots dP(x_{2\ell}). \end{aligned}$$

Next the inequality

$$\begin{aligned} P_{2\ell} &\geq \int_{R_\ell} \left\{ \int_{x_{\ell+1}} \dots \int_{x_{2\ell}} \theta \left[\sup_{\alpha \in \Lambda} S(x_1, \dots, x_{2\ell}; \alpha) - \frac{\varepsilon_0}{3} \right] \right. \\ &\quad \times dP(x_{\ell+1}) \dots dP(x_{2\ell}) \} dP(x_1) \dots dP(x_\ell). \end{aligned}$$

is valid. To each point x_1, \dots, x_ℓ belonging to R_ℓ we assign the value $\alpha^*(x_1, \dots, x_\ell)$ such that

$$\left| \frac{1}{\ell} \sum_{i=1}^{\ell} F(x_i, \alpha^*) - EF(x, \alpha^*) \right| > \varepsilon.$$

Denote by \hat{R}_ℓ the event in $X^\ell = \{x_{\ell+1}, \dots, x_{2\ell}\}$ such that

$$\left| \frac{1}{\ell} \sum_{i=\ell+1}^{2\ell} F(x_i, \alpha^*) - EF(x, \alpha^*) \right| \leq \frac{\varepsilon}{2}.$$

Furthermore,

$$\begin{aligned} P_{2\ell} &\geq \int_{R_\ell} \left\{ \int_{\hat{R}_\ell} \theta \left[S(x_1, \dots, x_{2\ell}; \alpha^*(x_1, \dots, x_\ell)) - \frac{\varepsilon}{2} \right] dP(x_{\ell+1}) \dots dP(x_{2\ell}) \right\} \\ &\quad dP(x_1) \dots dP(x_{2\ell}). \end{aligned}$$

However, if $(x_1, \dots, x_\ell) \in R_\ell$, while $(x_{\ell+1}, \dots, x_{2\ell}) \in \hat{R}_\ell$, then the integrand equals one. For $\ell > \frac{2}{\varepsilon}$ using Hoeffding inequality we obtain $P(\hat{R}_\ell) > \frac{1}{2}$ and therefore

$$P_{2\ell} > \frac{1}{2} \int_{R_\ell} dP(x_1) \dots dP(x_\ell) = \frac{1}{2} P(R_\ell).$$

This proves first part of the lemma.

To prove second part of the lemma, let us assume that there exists such $\varepsilon_0 > 0$ that $P(R_\ell) \neq 0$. Then according to (15.32) we have

$$\lim_{\ell \rightarrow \infty} P_{2\ell} \neq 0,$$

which contradicts the condition of the lemma. The lemma is thus proved.

The Proof of the Sufficiency of Conditions of the Theorem. We shall prove that under the conditions of the theorem we have

$$P \left\{ \sup_{\alpha \in \Lambda} S(x_1, \dots, x_{2\ell}; \alpha) > \varepsilon \right\} \xrightarrow{\ell \rightarrow \infty} 0.$$

In view of Lemma 15.5, it follows from conditions (15.33) that the statement of the theorem is fulfilled; that is,

$$P \left\{ \sup_{\alpha \in \Lambda} \left| \frac{1}{\ell} \sum_{i=1}^{\ell} F(x_i, \alpha) - EF(x, \alpha) \right| > \varepsilon \right\} \xrightarrow{\ell \rightarrow \infty} 0.$$

We shall show the validity of (15.33).

For this purpose we note that in view of symmetry of the definition of the measure, the equation

$$\begin{aligned} P \left\{ \sup_{\alpha \in \Lambda} S(x_1, \dots, x_{2\ell}, \alpha) > \varepsilon \right\} \\ = \frac{1}{(2\ell)!} \sum_{j=1}^{(2\ell)!} P \left\{ \sup_{\alpha \in \Lambda} S(T_j(x_1, \dots, x_{2\ell}), \alpha) > \varepsilon \right\} \\ = \int \left\{ \frac{1}{(2\ell)!} \sum_{j=1}^{(2\ell)!} \theta [\sup_{\alpha \in \Lambda} S(T_j(x_1, \dots, x_{2\ell}), \alpha) - \varepsilon] \right\} dP(x_1) \dots dP(x_{2\ell}) \end{aligned} \quad (15.34)$$

holds true. Here T_j , $j = 1, \dots, (2\ell)!$ are all the permutations of the indices, and $T_j(x_1, \dots, x_{2\ell})$ is a sequence of arguments obtained from the sequence $x_1, \dots, x_{2\ell}$ when the permutation T_j is applied.

Now consider the integrand in (15.34):

$$K = \frac{1}{(2\ell)!} \sum_{j=1}^{(2\ell)!} \theta [\sup_{\alpha \in \Lambda} S(T_j(x_1, \dots, x_{2\ell}), \alpha) - \varepsilon]$$

Let A be the set of vectors in $R_{2\ell}$ with coordinates $z^i = F(x_i, \alpha)$, $i = 1, \dots, 2\ell$ for all $\alpha \in \Lambda$.

Let $z(1), \dots, z(N_0)$ be the minimal proper ε net in A , and let $\alpha(1), \dots, \alpha(N_0)$ be the values of α such that

$$z^i(k) = F(x_i, \alpha(k)), \quad i = 1, \dots, 2\ell, \quad k = 1, \dots, N_0.$$

We show that if the inequality

$$\max_{1 \leq k \leq N_0} S(x_1, \dots, x_{2\ell}; \alpha(k)) < \frac{\varepsilon}{3}$$

is fulfilled, then the inequality

$$\sup_{\alpha \in \Lambda} S(x_1, \dots, x_{2\ell}, \alpha) < \varepsilon$$

is also valid.

Indeed, for any α there exists $\alpha(k)$ such that,

$$|F(x_i, \alpha) - F(x_i, \alpha(k))| < \frac{\varepsilon}{3}, \quad i = 1, \dots, 2\ell.$$

Therefore

$$\begin{aligned} & \left| \frac{1}{\ell} \sum_{i=1}^{\ell} F(x_i, \alpha) - \frac{1}{\ell} \sum_{i=\ell+1}^{2\ell} F(x_i, \alpha) \right| \\ &= \left| \frac{1}{\ell} \left(\sum_{i=1}^{\ell} F(x_i, \alpha) - \sum_{i=1}^{\ell} F(x_i, \alpha(k)) \right) \right. \\ &\quad \left. - \frac{1}{\ell} \left(\sum_{i=\ell+1}^{2\ell} F(x_i, \alpha) - \sum_{i=\ell+1}^{2\ell} F(x_i, \alpha(k)) \right) \right. \\ &\quad \left. + \frac{1}{\ell} \left(\sum_{i=1}^{\ell} F(x_i, \alpha(k)) - \sum_{i=\ell+1}^{2\ell} F(x_i, \alpha(k)) \right) \right| \\ &\leq 2 \frac{\varepsilon}{3} + \frac{1}{\ell} \left| \sum_{i=1}^{\ell} F(x_i, \alpha(k)) - \sum_{i=\ell+1}^{2\ell} F(x_i, \alpha(k)) \right| < \varepsilon. \end{aligned}$$

Analogous bounds are valid for $S(T_j(x_1, \dots, x_{2\ell}), \alpha)$. Therefore

$$\begin{aligned} K &= \frac{1}{(2\ell)!} \sum_{j=1}^{(2\ell)!} \theta \left[\max_k S(T_j(x_1, \dots, x_{2\ell}), \alpha(k)) - \frac{\varepsilon}{3} \right] \\ &\leq \frac{1}{(2\ell)!} \sum_{j=1}^{(2\ell)!} \sum_{k=1}^{N_0} \theta \left[S(T_j(x_1, \dots, x_{2\ell}), \alpha(k)) - \frac{\varepsilon}{3} \right] \\ &= \sum_{k=1}^{N_0} \left\{ \frac{1}{(2\ell)!} \sum_{j=1}^{(2\ell)!} \theta \left[S(T_j(x_1, \dots, x_{2\ell}), \alpha(k)) - \frac{\varepsilon}{3} \right] \right\}. \end{aligned}$$

We evaluate the expression in curly brackets:

$$I_1 = \frac{1}{(2\ell)!} \sum_{j=1}^{(2\ell)!} \theta \left[\frac{1}{\ell} \left| \sum_{i=1}^{\ell} F(x_{T_j(i)}, \alpha(k)) - \sum_{i=\ell+1}^{2\ell} F(x_{T_j(i)}, \alpha(k)) \right| - \frac{\varepsilon}{3} \right],$$

where $T_j(i)$ is the index into which the index i is transformed in the permutation T_j . We order the values $F(x_i, \alpha(k))$ by magnitude:

$$F(x_{i_1}, \alpha(k)) \leq \dots \leq F(x_{i_p}, \alpha(k))$$

and denote $z^p = F(x_{i_p}, \alpha(k))$.

Next we use the notations

$$\begin{aligned} \Delta_1 &= z^1, & \Delta_p &= z^p - z^{p-1}, \\ \delta_{i_p} &= \begin{cases} 1 & \text{for } F(x_i, \alpha(k)) \leq z^p, \\ 0 & \text{for } F(x_i, \alpha(k)) > z^p, \end{cases} \\ r_i^j &= \begin{cases} 1 & \text{for } T_j^{-1}(i) \leq \ell, \\ 0 & \text{for } T_j^{-1}(i) > \ell, \end{cases} \end{aligned}$$

where $T_j^{-1}(i)$ is the index which is transformed into i by the permutation T_j . Then

$$\begin{aligned} &\frac{1}{\ell} \left| \sum_{i=1}^{\ell} F(x_{T_j(i)}, \alpha(k)) - \sum_{i=\ell+1}^{2\ell} F(x_{T_j(i)}, \alpha(k)) \right| \\ &= \frac{1}{\ell} \left| \sum_p \Delta_p \sum_{i=1}^{2\ell} \delta_{i_p} r_i^j - \sum_p \Delta_p \sum_{i=1}^{2\ell} \delta_{i_p} (1 - r_i^j) \right| \\ &= \sum_p \Delta_p \left| \frac{1}{\ell} \sum_{i=1}^{2\ell} \delta_{i_p} (2r_i^j - 1) \right|. \end{aligned}$$

Furthermore, if the equality

$$\max_p \frac{1}{\ell} \left| \sum_{i=1}^{2\ell} \delta_{i_p} (2r_i^j - 1) \right| < \frac{\varepsilon}{3} \quad (15.35)$$

is fulfilled, then the inequality

$$\sum_p \Delta_p \frac{1}{\ell} \left| \sum_{i=1}^{2\ell} \delta_{i_p} (2r_i^j - 1) \right| < \frac{\varepsilon}{3} \sum_p \Delta_p \leq \frac{\varepsilon}{3} \quad (15.36)$$

is also valid. The condition (15.35) is equivalent to the following:

$$\max_p \theta \left[\frac{1}{\ell} \left| \sum_{i=1}^{2\ell} \delta_{i_p} (2r_i^j - 1) \right| - \frac{\varepsilon}{3} \right] = 0.$$

Thus we obtain

$$\begin{aligned} I_1 &< \frac{1}{(2\ell)!} \sum_{j=1}^{(2\ell)!} \max_p \theta \left[\frac{1}{\ell} \left| \sum_{i=1}^{2\ell} \delta_{ip} (2r_i^j - 1) \right| - \frac{\varepsilon}{3} \right] \\ &\leq \sum_p \left\{ \frac{1}{(2\ell)!} \sum_{j=1}^{(2\ell)!} \theta \left[\frac{1}{\ell} \left| \sum_{i=1}^{2\ell} \delta_{ip} (2r_i^j - 1) \right| - \frac{\varepsilon}{3} \right] \right\}. \end{aligned} \quad (15.37)$$

Let there be 2ℓ balls, of which

$$\sum_{i=1}^{2\ell} \delta_{ip} = m$$

are black, in an urn model without replacement. We select ℓ balls (without replacement). Then the expression in the curly brackets of (15.37) is the probability that the number of black balls chosen from the urn will differ from the number of remaining black balls by at least $\varepsilon\ell/3$. This value equals

$$\Gamma = \sum_k \frac{C_m^k C_{2\ell-m}^{\ell-k}}{C_{2\ell}^{\ell}},$$

where k runs over all the values such that

$$\left| \frac{k}{\ell} - \frac{m-k}{\ell} \right| > \frac{\varepsilon}{3}$$

In Chapter 4, Section 4.13 the quantity Γ is bounded as

$$\Gamma < 3 \exp \left\{ -\frac{\varepsilon^2 \ell}{9} \right\}.$$

Thus

$$I_1 < \sum_{p=1}^{2\ell} 3 \exp \left\{ -\frac{\varepsilon^2 \ell}{9} \right\} = 6\ell \exp \left\{ -\frac{\varepsilon^2 \ell}{9} \right\}.$$

Returning to the estimate of K , we obtain

$$K < 6\ell N_0^\Lambda \left(x_1, \dots, x_{2\ell}, \frac{\varepsilon}{3} \right) \exp \left\{ -\frac{\varepsilon^2 \ell}{9} \right\}. \quad (15.38)$$

Finally, for any $c > 0$ we obtain

$$\begin{aligned}
 & P \left\{ \sup_{\alpha \in \Lambda} \left| \frac{1}{\ell} \sum_{i=1}^{\ell} F(x_i, \alpha) - \frac{1}{\ell} \sum_{i=\ell+1}^{2\ell} F(x_i, \alpha) \right| > \varepsilon \right\} \\
 & \leq \int_{\ln N_0^{\Lambda}(x_1, \dots, x_{2\ell}; \varepsilon/3) > c\ell} dP(x_1) \dots dP(x_{2\ell}) \\
 & \quad + \int_{\ln N_0^{\Lambda}(x_1, \dots, x_{2\ell}; \varepsilon/3) \leq c\ell} K(x_1, \dots, x_{2\ell}) dP(x_1) \dots dP(x_{2\ell}) \\
 & \leq P \left\{ \frac{\ln N_0^{\Lambda}(x_1, \dots, x_{2\ell}; \varepsilon/3)}{\ell} > c \right\} + 6\ell \exp \left\{ -\frac{\varepsilon^2 \ell}{9} + c\ell \right\}.
 \end{aligned} \tag{15.39}$$

Setting $c < \varepsilon^2/10$, we obtain that the second term on the right-hand side approaches zero as ℓ increases. In view of the condition of the theorem and the relation (15.8), the first term tends to zero.

Thus, the theorem has been proven.

15.7 COROLLARIES FROM THEOREM 15.1

Theorem 15.2. *The inequality*

$$\begin{aligned}
 & P \left\{ \sup_{\alpha \in \Lambda} \left| EF(x, \alpha) - \frac{1}{\ell} \sum_{i=1}^{\ell} F(x_i, \alpha) \right| > \varepsilon \right\} \\
 & \leq 12\ell EN^{\Lambda}(x_1, \dots, x_{2\ell}; \varepsilon) \exp \left\{ -\frac{\varepsilon^2 \ell}{36} + c\ell \right\}
 \end{aligned}$$

holds true. The bound is nontrivial if

$$\lim_{\ell \rightarrow \infty} \frac{\ln EN^{\Lambda}(x_1, \dots, x_{2\ell}; \varepsilon)}{\ell} = 0.$$

To prove this theorem it is sufficient to put (15.38) into (15.34) and use the result of Lemma 15.5 for $\ell \geq \frac{2}{\varepsilon^2}$. For $\ell < \frac{2}{\varepsilon^2}$, the bound is trivial.

Theorem 15.3. *For uniform convergence of means to their mathematical expectations it is necessary and sufficient that for any $\varepsilon > 0$ the equality*

$$\lim_{\ell \rightarrow \infty} \frac{1}{\ell} E \ln V(A_\varepsilon) = \ln \varepsilon$$

be fulfilled, where A_ε is the ε extension of the set A .

Proof of Necessity. Let $\varepsilon, \delta > 0$ and $\delta < \varepsilon$ and let T_0 be a minimal proper δ net of A with the number of elements $N_0^\Lambda(x_1, \dots, x_{2\ell}; \delta)$. We assign to each point in T_0 a cube with an edge $\varepsilon + 2\delta$ and center at this point, oriented along the coordinate axes.

The union of these cubes contains A_ε and hence

$$V(A_\varepsilon) < N_0^\Lambda(x_1, \dots, x_{2\ell}; \delta)(\varepsilon + 2\delta)^\ell,$$

from which we obtain

$$\frac{1}{\ell} E \ln V(A_\varepsilon) \leq \frac{H^\Lambda(\delta, \ell)}{\ell} + \ln(\varepsilon + 2\delta).$$

In view of the basic theorem we obtain

$$\lim_{\ell \rightarrow \infty} E \frac{1}{\ell} \ln V(A_\varepsilon) \leq \ln(\varepsilon + 2\delta).$$

Since $V(A_\varepsilon) > \varepsilon^\ell$ and δ is arbitrary, we arrive at the required assertion.

Proof of Sufficiency. Assume that the uniform convergence is not valid. Then for some $\varepsilon > 0$ we have

$$\lim_{\ell \rightarrow \infty} \frac{1}{2\ell} E \ln N_0^\Lambda(x_1, \dots, x_{2\ell}; 1.5\varepsilon) = \gamma > 0,$$

from which, in view of Lemma 15.2, we obtain

$$\lim_{\ell \rightarrow \infty} E \frac{\ln V(A_\varepsilon)}{\ell} \geq \gamma + \ln \varepsilon$$

The theorem is proved.

Denote by $L(x_1, \dots, x_\ell; \varepsilon)$ the number of elements in a minimal ε net of the set $A(x_1, \dots, x_\ell)$ in the metric

$$\rho(z_1, z_2) = \frac{1}{\ell} \sum_{i=1}^{\ell} |z_1^i - z_2^i|.$$

Theorem 15.4. *For a uniform convergence of means to the mathematical expectations it is necessary and sufficient that a function $T(\varepsilon)$ exists such that*

$$\lim_{\ell \rightarrow \infty} P\{L(x_1, \dots, x_\ell; \varepsilon) > T(\varepsilon)\} = 0.$$

To prove this theorem we prove two lemmas.

Lemma 15.6. *If uniform convergence is valid in the class of functions $F(x, \alpha)$, $\alpha \in A$, then it is also valid in the class $|F(x, \alpha)|$, $\alpha \in A$.*

Proof The mapping

$$F(x, \alpha) \longrightarrow |F(x, \alpha)|$$

does not increase the distance

$$\rho(\alpha_1, \alpha_2) = \max_{1 \leq i \leq \ell} |F(x_i, \alpha_1) - F(x_i, \alpha_2)|$$

Therefore

$$N_0^\Lambda(x_1, \dots, x_\ell; \varepsilon) > \hat{N}_0^\Lambda(x_1, \dots, x_\ell; \varepsilon),$$

where N_0^Λ and \hat{N}_0^Λ are the minimal numbers of the elements in an ε net in the sets \mathbf{A} and $\hat{\mathbf{A}}$ representatively generated by the classes $F(x, \alpha)$ and $|F(x, \alpha)|$.

Consequently the condition

$$\lim_{\ell \rightarrow \infty} P \left\{ \frac{\ln N_0^\Lambda(x_1, \dots, x_\ell; \varepsilon)}{\ell} > \delta \right\} = 0$$

implies

$$\lim_{\ell \rightarrow \infty} P \left\{ \frac{\ln \hat{N}_0^\Lambda(x_1, \dots, x_\ell; \varepsilon)}{\ell} > \delta \right\} = 0.$$

The lemma is proved.

Consider a two-parameter class of functions

$$f(x, \alpha_1, \alpha_2) = |F(x, \alpha_1) - F(x, \alpha_2)|, \quad \alpha_1, \alpha_2 \in \Lambda,$$

along with the class of functions $F(x, \alpha)$, $\alpha \in A$.

Lemma 15.7. *Uniform convergence in the class $F(x, \alpha)$ implies uniform convergence in $f(x, \alpha_1, \alpha_2)$.*

Proof. Uniform convergence in $F(x, \alpha)$ clearly implies such a convergence in $|F(x, \alpha_1) - F(x, \alpha_2)|$. Indeed, the condition

$$\sup_{\alpha \in \Lambda} \left| E F(x, \alpha) - \frac{1}{\ell} \sum_{i=1}^{\ell} F(x_i, \alpha) \right| < \varepsilon$$

and the condition

$$\begin{aligned} & \left| EF(x, \alpha_1) - EF(x, \alpha_2) - \frac{1}{\ell} \sum_{i=1}^{\ell} F(x_i, \alpha_1) + \frac{1}{\ell} \sum_{i=1}^{\ell} F(x_i, \alpha_2) \right| \\ & \leq \left| EF(x, \alpha_1) - \frac{1}{\ell} \sum_{i=1}^{\ell} F(x_i, \alpha_1) \right| + \left| EF(x, \alpha_2) - \frac{1}{\ell} \sum_{i=1}^{\ell} F(x_i, \alpha_2) \right| \end{aligned}$$

imply that

$$\sup_{\alpha_1, \alpha_2 \in \Lambda} \left| E(F(x, \alpha_1) - F(x, \alpha_2)) - \frac{1}{\ell} \sum_{i=1}^{\ell} (F(x_i, \alpha_1) - F(x_i, \alpha_2)) \right| \leq 2\varepsilon.$$

Applying Lemma 15.6, we obtain the required result.

Proof of Theorem 15.4. To prove the necessity, note that according to Lemma 15.7 the uniform convergence of the class $F(x, \alpha)$ implies the uniform convergence of the class $f(x, \alpha_1, \alpha_2)$, that is,

$$\sup_{\alpha_1, \alpha_2 \in \Lambda} \left| E|F(x, \alpha_1) - F(x, \alpha_2)| - \frac{1}{\ell} \sum_{i=1}^{\ell} |F(x_i, \alpha_1) - F(x_i, \alpha_2)| \right| \xrightarrow[\ell \rightarrow \infty]{P} 0. \quad (15.40)$$

Consequently for any $\varepsilon > 0$ there exist finite ℓ_0 and sequence $x_1^*, \dots, x_{\ell_0}^*$ such that the left-hand side of (15.40) is smaller than ε . This means that the distance

$$\hat{\rho}_1(\alpha_1, \alpha_2) = \frac{1}{\ell_0} \sum_{i=1}^{\ell_0} |F(x_i^*, \alpha_1) - F(x_i^*, \alpha_2)| \quad (15.41)$$

approximates with accuracy ε the distance in the space $L_1(\mathbf{P})$

$$\hat{\rho}_2(\alpha_1, \alpha_2) = \int |F(x, \alpha_1) - F(x, \alpha_2)| dP(x) \quad (15.42)$$

uniformly in α_1 and α_2 . However, in the metric (15.41) there exists on the set $F(x, \alpha)$, $\alpha \in \Lambda$, a finite ε net S with the number of elements $L(x_1^*, \dots, x_{\ell_0}^*; \varepsilon)$. The same net S forms a $2s$ net in the space Λ with the metric (15.42). Next we utilize the uniform convergence of $\hat{\rho}_1(\alpha_1, \alpha_2)$ to $\hat{\rho}_2(\alpha_1, \alpha_2)$ and obtain that the same net S , with probability tending to one as $\ell \rightarrow \infty$, forms a $3s$ net on the set $\Lambda(x_1^*, \dots, x_{\ell_0}^*)$. Setting

$$T(\varepsilon) = L(x_1^*, \dots, x_{\ell_0}^*; \varepsilon),$$

we obtain the assertion of the theorem.

The proof of sufficiency of the conditions is analogous to the proof of sufficiency for Theorem 15.1.

16

NECESSARY AND SUFFICIENT CONDITIONS FOR UNIFORM ONE-SIDED CONVERGENCE OF MEANS TO THEIR EXPECTATIONS

In this chapter we achieve our main goal: We derive the necessary and sufficient conditions of uniform one-sided convergence of the means to their expectations over a given set of bounded functions $F(x, \alpha)$, $\alpha \in A$; in other words, we derive the conditions under which the limit

$$\lim_{\ell \rightarrow \infty} P \left\{ \sup_{\alpha \in \Lambda} \left(EF(x, \alpha) - \frac{1}{\ell} \sum_{i=1}^{\ell} F(x_i, \alpha) \right) > \varepsilon \right\} = 0$$

holds true for any $\varepsilon > 0$.

16.1 INTRODUCTION

In Chapter 15 we discussed the problem of uniform two-sided convergence of the means to their mathematical expectations over the set of functions $F(x, \alpha)$, $\alpha \in A$:

$$\lim_{\ell \rightarrow \infty} P \left\{ \sup_{\alpha \in \Lambda} \left| EF(x, \alpha) - \frac{1}{\ell} \sum_{i=1}^{\ell} F(x_i, \alpha) \right| > \varepsilon \right\} = 0.$$

The existence of uniform two-sided convergence forms the sufficient conditions for consistency of the empirical risk minimization induction principle.

However, uniform two-sided convergence is too strong a requirement for justification of the principle of the empirical risk minimization. In Chapter 3

we proved that for consistency of the empirical risk minimization principle, it is necessary and sufficient that the uniform one-sided convergence

$$\lim_{\ell \rightarrow \infty} P \left\{ \sup_{\alpha \in \Lambda} \left(EF(x, \alpha) - \frac{1}{\ell} \sum_{i=1}^{\ell} F(x_i, \alpha) \right) > \varepsilon \right\} = 0$$

be true.

In this chapter we will derive the necessary and sufficient conditions for uniform one-sided convergence.

In Chapter 3 we gave a special definition for nontrivial consistency of the maximum likelihood method (which requires the consistency of estimating any density from a given set of densities). In this chapter we prove that the necessary and sufficient conditions imply the (nontrivial) consistency for the maximum likelihood method as well.

To derive the necessary and sufficient conditions of uniform one-sided convergence, we have to consider several constructions.

16.2 MAXIMUM VOLUME SECTIONS

Let there be specified a space X , a probability measure $P(x)$, and a set of functions

$$a \leq F(x, a) \leq A, \quad a \in A$$

measurable with respect to $P(x)$. (Without restrictions in generality we assume that $a = 0$ and $A = 1$.)

Let

$$x_1, \dots, x_\ell$$

be an i.i.d. sample from X .

We consider an ℓ -ary space and a set Z of vectors $z = (z^1, \dots, z^\ell)$ defined on it by the rule

$$Z = \left\{ z : \exists \alpha \in \Lambda, \forall i z^i = F(x_i, \alpha) \right\}.$$

We also consider an ε extension of the set Z —that is, a set of ℓ -ary vectors

$$Y_\varepsilon = \left\{ y : \exists z \in Z, \rho(y, z) < \frac{\varepsilon}{2} \right\},$$

$$\rho(y, z) = \max_{1 \leq i \leq \ell} |y^i - z^i|.$$

Here Y_ε is the union of all open cubes oriented along the coordinate axes, with edge ε and with center at the points of the set Z .

Let $V_\varepsilon(x_1, \dots, x_\ell)$ be a volume of the set $Y_\varepsilon = Y_\varepsilon(x_1, \dots, x_\ell)$. We denote

$$H_\varepsilon^\Lambda(\ell) = E \ln V_\varepsilon(x_1, \dots, x_\ell). \quad (16.1)$$

In Theorem 15.1 (Chapter 15, Section 15.1) it is shown that the conditions

$$\lim_{\ell \rightarrow \infty} \frac{H_e^\Lambda(\varepsilon, \ell)}{\ell} = 0, \quad \forall \varepsilon > 0 \quad (16.2)$$

form the necessary and sufficient conditions for uniform two-sided convergence and are equivalent to

$$\lim_{\ell \rightarrow \infty} \frac{H_e^\Lambda(\ell)}{\ell} - C_e^\Lambda = \ln \varepsilon. \quad (16.3)$$

Now suppose that the conditions defined by (16.3) are not satisfied; that is, the conditions for uniform two-sided convergence are equivalent to

$$\lim_{\ell \rightarrow \infty} \frac{H_e^\Lambda(\ell)}{\ell} = C_e^\Lambda > \ln \varepsilon. \quad (16.4)$$

We set $x = x_1$ and construct the set $Y_{\varepsilon, b}(x, x_2, \dots, x_\ell)$ for the specified x_2, \dots, x_ℓ . Consider the section produced on cutting this set by a hyperplane

$$y^1 = b$$

as a set $Y_{\varepsilon, b}(x, x_2, \dots, x_\ell)$ of vectors $y = (y^2, \dots, y^\ell) \in R^{\ell-1}$ such that there exists $a^* \in A$ satisfying the conditions

$$\begin{aligned} |F(x, a^*) - b| &< \frac{\varepsilon}{2}, \\ |F(x_i, a^*) - y^i| &< \frac{\varepsilon}{2}, \quad i = 2, \dots, \ell. \end{aligned}$$

The volume $V_{\varepsilon, b}(x, x_2, \dots, x_\ell)$ of the set $Y_{\varepsilon, b}(x, x_2, \dots, x_\ell)$ is nonzero for b 's such that

$$\exists a \in A : |F(x, a) - b| < \frac{\varepsilon}{2}. \quad (16.5)$$

Obviously, if (16.5) is satisfied, then the inequalities

$$\varepsilon^{\ell-1} \leq V_{\varepsilon, b}(x, x_2, \dots, x_\ell) \leq (1 + \varepsilon)^{\ell-1}$$

hold true. Accordingly, the inequalities

$$(\ell - 1) \ln \varepsilon \leq \ln V_{\varepsilon, b}(x, x_2, \dots, x_\ell) \leq (\ell - 1) \ln(1 + \varepsilon)$$

hold true. We denote

$$\begin{aligned} H_{\varepsilon, b}^\Lambda(x, \ell) &= E_{x_2, \dots, x_\ell} \ln V_{\varepsilon, b}(x, x_2, \dots, x_\ell), \\ C_{\varepsilon, b}^\Lambda(x) &= \lim_{\ell \rightarrow \infty} \frac{H_{\varepsilon, b}^\Lambda(x, \ell)}{\ell} \end{aligned}$$

The functions $H_{\varepsilon, b}^{\Lambda}(x, \ell)$ and $C_{\varepsilon, b}^{\Lambda}(x)$ are defined on the set of b 's satisfying **(16.4)**. In the domain of definition for $C_{\varepsilon, b}^{\Lambda}(x)$ the following inequalities hold true:

$$\ln \varepsilon \leq C_{\varepsilon, b}^{\Lambda}(x) \leq \ln(1 + \varepsilon).$$

We denote by D the set of b 's such that

$$C_{\varepsilon}^{\Lambda}(x) = \max_b C_{\varepsilon, b}^{\Lambda}(x).$$

The following theorem is valid:

Theorem 16.1. *For almost all x 's,*

$$C_{\varepsilon}^{\Lambda}(x) = C_{\varepsilon}^{\Lambda}. \quad (16.6)$$

Equality (16.6) is satisfied on the set D_x^{Λ} formed by an aggregation of a finite number of intervals of a length each not less than ε .

To prove the Theorem **16.1**, we consider the set $Y_{\varepsilon, \delta, b}(x_1, \dots, x_\ell) \subset R^{\ell-1}$, $\varepsilon > 0$, $\delta > 0$, of vectors $y = (y^2, \dots, y^\ell)$ such that there exists $\alpha \in A$ for which

$$(a) \quad |F(x_1, \alpha) - b| < \frac{\delta}{2}$$

and

$$(b) \quad |F(x_i, \alpha) - y^i| < \frac{\varepsilon}{2}, \quad i = 2, \dots, \ell.$$

Accordingly we denote by $V_{\varepsilon, \delta, b}(x_1, \dots, x_\ell)$ the volume of the set $Y_{\varepsilon, \delta, b}(x_1, \dots, x_\ell)$:

$$H_{\varepsilon, \delta, b}^{\Lambda}(x, \ell) = E_{x_2, \dots, x_\ell} \ln V_{\varepsilon, \delta, b}(x_1, \dots, x_\ell),$$

$$C_{\varepsilon, \delta, b}^{\Lambda}(x) = \lim_{\ell \rightarrow \infty} \frac{H_{\varepsilon, \delta, b}^{\Lambda}(x, \ell)}{\ell}$$

Here the functions $H_{\varepsilon, \delta, b}^{\Lambda}(x, \ell)$ and $C_{\varepsilon, \delta, b}^{\Lambda}(x)$ are defined for b 's such that

$$\exists \alpha: \quad |F(x, \alpha) - b| < \frac{\delta}{2}.$$

Obviously, the following equality holds

$$C_{\varepsilon, b}(x) = C_{\varepsilon, \varepsilon, b}(x).$$

Lemma 16.1. For almost all x 's,

$$\max_b C_{\varepsilon, \delta, b}^{\Lambda}(x) = C_{\varepsilon}^{\Lambda},$$

where the maximum is taken in the domain of definition for $C_{\varepsilon, \delta, b}^{\Lambda}(x)$.

Proof. (a) First we show that the inequality

$$\sup_b C_{\varepsilon, \delta, b}^{\Lambda}(x) \leq C_{\varepsilon}^{\Lambda} \quad (16.7)$$

holds true. Indeed

$$Y_{\varepsilon, \delta, b}(x, x_2, \dots, x_\ell) \subset Y_{\varepsilon}(x_2, \dots, x_\ell)$$

because the definition of the set $Y_{\varepsilon, \delta, b}(x_1, x_2, \dots, x_\ell)$ differs from that of the set $Y_{\varepsilon}(x_2, \dots, x_\ell)$ in only the additional condition (a).

Therefore

$$V_{\varepsilon, \delta, b}(x, x_2, \dots, x_\ell) \leq V_{\varepsilon}(x_2, \dots, x_\ell)$$

from which (16.7) follows immediately.

(b) Now we will show that for almost all x 's there exists b such that

$$C_{\varepsilon, \delta, b}^{\Lambda}(x) \geq C_{\varepsilon}^{\Lambda}. \quad (16.8)$$

To this end, we consider a finite sequence of numbers b_0, b_1, \dots, b_k such that

$$\begin{aligned} b_0 &= 0, \\ b_k &= 1, \\ b_{i+1} - b_i &< \frac{\delta}{2}, \quad i = 0, 1, \dots, k-1. \end{aligned}$$

The number k depends solely on δ .

Note that the equality

$$\bigcup_{i=0}^k Y_{\varepsilon, \delta, b_i}(x, x_2, \dots, x_\ell) = Y_{\varepsilon}(x_2, \dots, x_\ell)$$

holds true (because the union of δ -neighborhoods of the points b_i covers the entire set of values of $F(x, a)$). Therefore

$$\begin{aligned} V_{\varepsilon}(x_2, \dots, x_\ell) &\leq \sum_{i=0}^k V_{\varepsilon, \delta, b_i}(x, x_2, \dots, x_\ell) \\ &\leq (k+1) \max_i V_{\varepsilon, \delta, b_i}(x, x_2, \dots, x_\ell). \end{aligned}$$

Hence

$$\frac{\ln V_\varepsilon(x_2, \dots, x_\ell)}{\ell - 1} \leq \frac{\ln(k + 1)}{\ell - 1} + \max_i \frac{\ln V_{\varepsilon, \delta, b_i}(x, x_2, \dots, x_\ell)}{\ell - 1} \quad (16.9)$$

In Chapter 15 (Section 15.7) it is shown that

$$\frac{\ln V_\varepsilon(x_1, \dots, x_\ell)}{\ell} \xrightarrow[\ell \rightarrow \infty]{P} C_\varepsilon^\Lambda.$$

Similarly, it is established that

$$\frac{\ln V_{\varepsilon, \delta, b}(x, x_2, \dots, x_\ell)}{\ell - 1} \xrightarrow[\ell \rightarrow \infty]{P} C_{\varepsilon, \delta, b}^\Lambda(x) \quad (16.10)$$

in the domain of definition. Furthermore, since k is independent of ℓ , it follows from (16.8) and (16.9) that

$$\max_i C_{\varepsilon, \delta, b_i}^\Lambda(x) \geq C_\varepsilon^\Lambda. \quad (16.11)$$

The inequalities (16.8) and (16.11) prove the lemma.

Lemma 16.2. *Given $\delta < \varepsilon$, for almost all x 's there holds*

$$C_{\varepsilon, b}^\Lambda(x) = \max_{\beta_1 \leq b^* \leq \beta_2} C_{\varepsilon, \delta, b^*}^\Lambda(x),$$

where

$$\beta_1 = b - \frac{\varepsilon - \delta}{2}, \quad \beta_2 = b + \frac{\varepsilon - \delta}{2}.$$

Proof. By definition,

$$C_{\varepsilon, b}^\Lambda(x) = C_{\varepsilon, b}^{A^*},$$

where A^* is derived from A by imposing an additional constraint

$$|F(x, \alpha) - b| < \frac{\varepsilon}{2}. \quad (16.12)$$

By applying Lemma 16.1 to this subclass, we get

$$C_{\varepsilon, b}^{A^*} = \max_{\beta_1 \leq b^* \leq \beta_2} C_{\varepsilon, \delta, b^*}^\Lambda(x). \quad (16.13)$$

Moreover, in view of (16.12), we can strengthen Lemma 16.1 and obtain

$$C_{\varepsilon, b}^{A^*}(x) = \max_{\beta_1 \leq b^* \leq \beta_2} C_{\varepsilon, \delta, b^*}^{A^*}(x), \quad (16.14)$$

where β_1 and β_2 are defined in the condition of Lemma 16.2. Indeed, we can choose the sequence b_0, \dots, b_k used to prove Lemma 16.1 so that

$$b_0 = b - \frac{\varepsilon - \delta}{2}, \quad b_k = b + \frac{\varepsilon - \delta}{2},$$

and $b_{i+1} - b_i < \delta/2$ (in this case the union of neighborhoods covers the entire set of values of $F(x, a)$, $a \in A$). By repeating the proof of Lemma 16.1, we obtain (16.14). But for b from the segment $[\beta_1, \beta_2]$ the narrowing of A to A^* is immaterial because, given

$$|F(x, \alpha) - b| < \frac{\delta}{2},$$

the condition

$$|F(x, \alpha) - b| < \frac{\varepsilon}{2}$$

will be satisfied automatically. Thus

$$C_{\varepsilon, b}^\Lambda(x) = \max_{\beta_1 \leq b^* \leq \beta_2} C_{\varepsilon, \delta, b^*}^\Lambda(x). \quad (16.15)$$

This completes the proof of Lemma 16.2.

Proof of Theorem 16.1. The proof of the first part of Theorem 16.1 follows from Lemma 16.1 on inserting ε for 6.

The proof of the second part of the theorem follows from the fact that, by virtue of Lemma 16.2, for any point b where

$$C_{\varepsilon, b}^\Lambda(x) = C_\varepsilon^\Lambda$$

and for any $0 < \delta < \varepsilon$, one can find a point b^* such that

$$|b^* - b| < \frac{\varepsilon - \delta}{2} \quad (16.16)$$

and

$$C_{\varepsilon, \delta, b^*}^\Lambda = C_\varepsilon^\Lambda. \quad (16.17)$$

But then, by virtue of the same lemma and over the entire segment

$$b^* - \frac{\varepsilon - \delta}{2} \leq b \leq b^* + \frac{\varepsilon - \delta}{2}$$

the following equality will be satisfied:

$$C_{\varepsilon, b}^\Lambda(x) = C_\varepsilon^\Lambda$$

In view of the arbitrary smallness of 6, the conclusion of Theorem 16.1 follows from this.

16.3 THE THEOREM ON THE AVERAGE LOGARITHM

We denote by K_x^Λ the Lebesgue measure of the set D_x^Λ .

Theorem 16.2. *The inequality*

$$\ln K_x^\Lambda dP(x) \geq C_\varepsilon^\Lambda \quad (16.18)$$

holds true.

To prove Theorem 16.2, we fix x_1, \dots, x_ℓ and define the density $g(y)$ in \mathbb{R}^ℓ :

$$g(y) = \begin{cases} 0 & \text{if } y \notin Y_\varepsilon(x_1, \dots, x_\ell), \\ \frac{1}{V_\varepsilon(x_1, \dots, x_\ell)} & \text{if } y \in Y_\varepsilon(x_1, \dots, x_\ell). \end{cases}$$

On any coordinate subspace R^k this density induces the density

$$\begin{aligned} g_k(y) &= g_{R^k}(y^1, \dots, y^k; x_1, \dots, x_\ell) \\ &= \int_{-\varepsilon/2}^{(1+\varepsilon)/2} \cdots \int_{-\varepsilon/2}^{(1+\varepsilon)/2} g(y^1, \dots, y^\ell) dy^{k+1}, \dots, dy^\ell. \end{aligned}$$

We denote

$$\hat{H}_{R^k}(x_1, \dots, x_\ell) = - \int_{R^k} g_k(y) \ln g_k(y) dy,$$

where $H_{R^k}(\cdot)$ is the Shannon entropy of the density g_k . In particular,

$$\hat{H}_{R^1}(x_1, \dots, x_\ell) = - \int_{Y_\varepsilon(x_1, \dots, x_\ell)} \frac{1}{V_\varepsilon(x_1, \dots, x_\ell)} \ln \frac{1}{V_\varepsilon(x_1, \dots, x_\ell)} dy.$$

We denote

$$\hat{H}_{R^k}(\ell) = E_{x_1, \dots, x_\ell} \hat{H}_{R^k}(x_1, \dots, x_\ell).$$

Note that, with specified ℓ in advance, $H_{R^k}(\ell)$ depends solely on the dimensionality of k and is independent of the choice of a specific subspace (which follows from the independence of the sample): that is,

$$\hat{H}_{R^k}(\ell) = \hat{H}(k - \ell).$$

Therefore,

$$\hat{H}(\ell, \ell) = H_\varepsilon^\Lambda(\ell).$$

We denote

$$\Delta(k, \ell) = \begin{cases} \hat{H}(1, \ell) & \text{for } k = 1, \\ \hat{H}(k, \ell) - \hat{H}(k-1, \ell) & \text{for } k > 1. \end{cases}$$

Lemma 16.3. *The following inequality*

$$\Delta(k+1, \ell) \leq \Delta(k, \ell)$$

holds true.

Proof. Let R^k and R^{k+1} ($R^k \subset R^{k+1}$) be two coordinate subspaces with coordinates y^1, \dots, y^k and y^1, \dots, y^k, y^s , respectively. We denote

$$\Delta(R^{k+1}, R^k) = \hat{H}_{R^{k+1}}(x_1, \dots, x_\ell) - \hat{H}_{R^k}(x_1, \dots, x_\ell)$$

We have

$$\begin{aligned} \hat{H}_{R^{k+1}}(x_1, \dots, x_\ell) &= - \int_{R^{k+1}} g_{R^{k+1}}(y) \ln g_{R^{k+1}}(y) dy \\ &= - \int_{y^s} \left\{ \int_{R^k} g(y^s | y^1, \dots, y^k) g(y^1, \dots, y^k) \right. \\ &\quad \times \ln \left[g(y^s | y^1, \dots, y^k) g(y^1, \dots, y^k) \right] dy^1, \dots, dy^k \left. \right\} dy^s \\ &= \hat{H}_{R^k}(x_1, \dots, x_\ell) \\ &\quad - \int_{y^s} \left\{ \int_{R^k} g(y^s | y^1, \dots, y^k) g(y^1, \dots, y^k) \ln \left[g(y^s | y^1, \dots, y^k) \right] dy^1, \dots, dy^k \right\} dy^s. \end{aligned}$$

Hence

$$\begin{aligned} \Delta(R^{k+1}, R^k) &= \int_{y^s} \left\{ \int_{R^k} g(y^s | y^1, \dots, y^k) g(y^1, \dots, y^k) \ln \left[g(y^s | y^1, \dots, y^k) \right] dy^1, \dots, dy^k \right\} dy^s; \end{aligned} \tag{16.19}$$

that is, $\Delta(R^{k+1}, R^k)$ is the average conditional entropy for the specified Y^1, \dots, Y^k .

We now put:

- R^k – the space y^1, \dots, y^k ;
- R^{k+1} – the space y^1, \dots, y^k, y^{k+1} ;
- R_1^{k+1} – the space y^1, \dots, y^k, y^{k+2} ;
- R^{k+2} – the space $y^1, \dots, y^k, y^{k+1}, y^{k+2}$.

By applying (16.19) to pairs $\Delta(R_1^{k+1}, R^k)$ and $\Delta(R^{k+2}, R^{k+1})$, respectively, and recalling the theorem of nonnegativity of information proved in information theory, we get

$$\Delta(R_1^{k+1}, R^k) \geq \Delta(R^{k+2}, R^{k+1}).$$

By averaging over the sample x_1, \dots, x_ℓ , we obtain

$$E_{x_1, \dots, x_\ell} \Delta(R_1^{k+1}, R^k) \geq E_{x_1, \dots, x_\ell} \Delta(R^{k+2}, R^{k+1}). \quad (16.20)$$

Hence

$$\Delta(k, \ell) \geq \Delta(k+1, \ell).$$

In accordance with the above note, Eq. (16.20), in contrast to (16.19), depends solely on the dimensionality and does not depend on the choice of specific subspace. Lemma 16.3 has thus been proved.

Corollary 1. *The inequality*

$$\hat{H}(1, \ell) \geq \frac{\hat{H}(\ell, \ell)}{\ell} = \frac{H_\varepsilon^\Lambda(\ell)}{\ell}$$

holds true.

Indeed, since

$$\begin{aligned} H_\varepsilon^\Lambda(\ell) &= \hat{H}(\ell, \ell) = \sum_{k=1}^{\ell} \Delta(k, \ell), \\ \Delta(1, \ell) &= \hat{H}(1, \ell) \end{aligned}$$

the validity of Corollary 16.1 stems from Lemma 16.3.

Corollary 2. *The inequality*

$$\lim_{\ell \rightarrow \infty} \hat{H}(1, \ell) \geq C_\varepsilon^\Lambda$$

holds true.

Lemma 16.4. *For almost all x's, the inequality*

$$C_{\varepsilon,y}^{\Lambda}(x) < C_{\varepsilon}^{\Lambda}$$

(i.e., the condition $y \in D_x^{\Lambda}$) entails

$$g_1(y; x_1, \dots, x_{\ell}) \xrightarrow[\ell \rightarrow \infty]{P} 0.$$

Proof. By definition,

$$g_1(y; x_1, \dots, x_{\ell}) = \frac{V_{\varepsilon,y}(x, \dots, x_{\ell})}{V_{\varepsilon}(x, \dots, x_{\ell})}$$

On the other hand, as was noted earlier,

$$\begin{aligned} \frac{\ln V_{\varepsilon}(x, \dots, x_{\ell})}{\ell} &\xrightarrow[\ell \rightarrow \infty]{P} C_{\varepsilon}^{\Lambda} \\ \frac{\ln V_{\varepsilon,y}(x, \dots, x_{\ell})}{\ell} &\xrightarrow[\ell \rightarrow \infty]{P} C_{\varepsilon,y}^{\Lambda}(x). \end{aligned}$$

Therefrom follows the conclusion of Lemma 16.4.

We denote further

$$\hat{H}_x(k, \ell) = \hat{H}_{R^k}(x, x_2, \dots, x_{\ell}),$$

where $R^k = \{y : y^1, \dots, y^k\}$.

Lemma 16.5. *For almost all x's and for all ε^* 's, the following relation holds true:*

$$P\{\hat{H}_{R^1}(x, x_2, \dots, x_{\ell}) > \ln K_x^{\Lambda} + \varepsilon^*\} \xrightarrow[\ell \rightarrow \infty]{P} 0.$$

Proof. Consider the 6 extension $D_x^{\Lambda}(\delta)$ of the set D_x^{Λ} —that is, the set of such y's for which there exists $y^* \in D_x^{\Lambda}$ satisfying the conditions

$$|y - y^*| < \frac{\delta}{2}, \quad \delta > 0.$$

Let $\bar{D}_x^{\Lambda}(\delta)$ be the complement to the δ extension of the $D_x^{\Lambda}(\delta)$ on the segment $-\varepsilon/2 \leq y \leq 1 + \varepsilon/2$. By Theorem 16.1, $D_x^{\Lambda}(\delta)$ consists of a finite number of closed intervals. We will now show that for all x's we have

$$\sup_{y \in \bar{D}_x^{\Lambda}(\delta)} g_1(y; x, x_2, \dots, x_{\ell}) \xrightarrow[\ell \rightarrow \infty]{P} 0. \quad (16.21)$$

Obviously, it will suffice to prove the above statement for an arbitrary segment Ω constituting $\bar{D}_x^\Lambda(\delta)$. Let

$$\Omega = [a_1, a_2], \quad a_1 = b_1 < b_2 < \dots < b_k = a_2, \quad b_{i+1} - b_i < \varepsilon,$$

where k is only function of ε .

We denote by $Y_{\varepsilon, \Omega}(x)$ the set of vectors $y = (y^1, \dots, y^\ell)$ such that there exists $a \in A$ satisfying the conditions

$$\begin{aligned} a_1 - \frac{\varepsilon}{2} &\leq F(x, \alpha) \leq a_2 + \frac{\varepsilon}{2}, \\ |F(x_i, \alpha) - y^i| &< \frac{\varepsilon}{2}, \quad i = 1, \dots, \ell. \end{aligned}$$

Then, by definition,

$$\begin{aligned} Y_{\varepsilon, y}(x) &\subset Y_{\varepsilon, \Omega}(x) \quad \text{for } y \in \Omega; \\ Y_{\varepsilon, \Omega}(x) &\subset \bigcup_{i=1}^k Y_{\varepsilon, b_i}(x). \end{aligned}$$

Hence

$$\sup_{y \in \Omega} V_{\varepsilon, y}(x) \leq V_{\varepsilon, \Omega}(x) \leq \sum_{i=1}^k V_{\varepsilon, b_i}(x)$$

and, as a consequence,

$$\sup_{y \in \Omega} g(y) = \sup_{y \in \Omega} \frac{V_{\varepsilon, y}(x)}{V_\varepsilon(x)} \leq \sum_{i=1}^k \frac{V_{\varepsilon, b_i}(x)}{V_\varepsilon(x)} \leq k \max_i g(b_i)$$

such that for all i 's, $b_i \in \Omega$ and

$$C_{\varepsilon, b_i}^\Lambda(x) < C_\varepsilon^\Lambda.$$

But, by Lemma 16.4,

$$\max_i g(b_i) \xrightarrow[\ell \rightarrow \infty]{P} 0,$$

from which we get (16.21). Furthermore, we have

$$\begin{aligned} \hat{H}_{R^1}(x, x_2, \dots, x_\ell) &= - \int_{-\varepsilon/2}^{1+\varepsilon/2} g(y) \ln g(y) dy \\ &= - \int_{y \in D_x^\Lambda(\delta)} g(y) \ln g(y) dy - \int_{y \in \bar{D}_x^\Lambda(\delta)} g(y) \ln g(y) dy \\ &\leq \ln \mathcal{L}(D_x^\Lambda(\delta)) + \mathcal{L}(\bar{D}_x^\Lambda(\delta)) \sup_{y \in D_x^\Lambda(\delta)} [g(y) \ln g(y)], \end{aligned}$$

where $\mathcal{L}(A)$ is Lebesgue measure of the set A .

Hence, by (16.21)' for all $\varepsilon^* > 0$ and for almost all x 's we have

$$P_{x,x_2,\dots,x_\ell} \left\{ \hat{H}_{R^1}(x, \dots, x_\ell) > \ln \mathcal{L}(D_x^\Lambda(\delta)) + \varepsilon^* \right\} \xrightarrow[\ell \rightarrow \infty]{} 0. \quad (16.22)$$

Furthermore, since (16.22) holds for any $\delta > 0$ and

$$\mathcal{L}(D_x^\Lambda(\delta)) \xrightarrow[\delta \rightarrow 0]{} K_x^\Lambda,$$

we find that for almost all x 's

$$P \left\{ \hat{H}_{R^1}(x, x_2, \dots, x_\ell) > \ln K_x^\Lambda + \varepsilon^* \right\} \xrightarrow[\ell \rightarrow \infty]{P} 0 \quad (16.23)$$

for all $\varepsilon^* > 0$. This completes the proof of Lemma 16.5.

Lemma 16.6. *Let $|F_\ell(x, y)| < B$ be a function of two variables, $x \in X$ and $y \in Y$ measurable on $X \times Y$ with $P(x, y) = P(x)P(y)$.*

Let there exist a function $\phi(x)$ measurable on X , such that for almost all x 's and for all $\varepsilon > 0$ we obtain

$$\lim_{\ell \rightarrow \infty} P_y \{ F_\ell(x, y) \geq \phi(x) + \varepsilon \} = 0.$$

Then

$$\overline{\lim}_{\ell \rightarrow \infty} E_{x,y} F_\ell(x, y) < \int \phi(x) dP(x).$$

Proof. We set $\varepsilon > 0$, $\delta > 0$. We denote by A_ℓ the event in the space X :

$$A_\ell = \{x : P_y \{ F_\ell(x, y) \geq \phi(x) + \varepsilon \} > \eta\}.$$

From the statement of Lemma 16.6 we have

$$\lim_{\ell \rightarrow \infty} P_x(A_\ell) = 0. \quad (16.24)$$

Furthermore,

$$\begin{aligned} E_{x,y} F_\ell(x, y) &= \int_X \int_Y F_\ell(x, y) dP(x) dP(y) \leq B P_x(A_\ell) \\ &\quad + \int_{x \in A_\ell} \left[\int_{-F_\ell(x,y) > \phi(x)+\varepsilon} F_\ell(x, y) dP(y) \right. \\ &\quad \left. + \int_{F_\ell(x,y) \leq \phi(x)+\varepsilon} F_\ell(x, y) dP(y) \right] dP(x) \\ &\leq B P_x(A_\ell) + B \eta + \int_X (\phi(x) + \varepsilon) dP(x). \end{aligned}$$

Taking into account (16.24) and the arbitrary smallness of η and ε , thus we have obtained the statement of Lemma 16.6.

Proof of Theorem 16.2. In Lemma 16.6, we put

$$x = x_1, \quad y = x_2, \dots, x_\ell,$$

$$F_\ell(x, y) = H(x_1, \dots, x_\ell) \leq \ln(1 + \varepsilon),$$

$$\phi(x) \leq K_x^\Lambda.$$

Subject to (16.23), we obtain

$$\overline{\lim}_{\ell \rightarrow \infty} H(1, \ell) \leq \int \ln K_x^\Lambda dP(x).$$

Combining the above result with that of Lemma 16.4, we get

$$C_\varepsilon^\Lambda \leq \liminf_{\ell \rightarrow \infty} H(1, \ell) \leq \overline{\lim}_{\ell \rightarrow \infty} H(1, \ell) \leq \int \ln K_x^\Lambda dP(x).$$

This completes the proof of Theorem 16.2.

16.4 THEOREM ON THE EXISTENCE OF A CORRIDOR

Definition. We call the corridor R_ε^Λ the set of pairs (x, D_x^Λ) .

Theorem 16.3. *To each point x we may let correspond the set D_x^Λ on the segment $[-\varepsilon, 1 + \varepsilon]$, with Lebesgue measure K_x^0 such that the following conditions are satisfied:*

1. $K_x^0 \geq \varepsilon;$
2. *For almost all x_1, \dots, x_ℓ (in the sense of the measure $P(x, \dots, x_\ell)$) and almost all y^1, \dots, y^ℓ , $y^i \in D_x^0$ (in the sense of Lebesgue measure), there can be found $\alpha^* \in A$ such that*

$$|F(x_i, \alpha^*) - y^i| \leq \frac{\varepsilon}{2}$$

for all $i = 1, \dots, \ell$.

To prove this theorem we consider the following process. Let us call it the process A.

Definition of the Process A. Let there be specified an infinite sample

$$x_1, \dots, x_\ell, \dots,$$

collected in a series of independent trials with distribution $P(x)$. At each step of the process, we will construct:

1. The subset Λ_i of functions $F(x, \alpha), \alpha \in A$, $\Lambda_i \subset A$.
2. The corridor $R_\varepsilon^{\Lambda_i}$ corresponding to this subset.

We put $\Lambda_1 = A$ and $R_\varepsilon^{\Lambda_1} = R_\varepsilon^\Lambda$.

Suppose that A , and $R_\varepsilon^{\Lambda_i}$ are constructed by the i th step, that is. $D_{x_i}^{\Lambda_i}$ is specified for each $x \in X$ and the point x_i occurs at that step. We choose the number $y^i \in D_{x_i}^{\Lambda_i}$ in a random fashion and independently in keeping with the density of the distribution

$$P(y) = \begin{cases} 0 & \text{if } y \notin D_{x_i}^{\Lambda_i}, \\ \frac{1}{K_{x_i}^{\Lambda_i}} & \text{if } y \in D_{x_i}^{\Lambda_i}. \end{cases}$$

We set

$$\Lambda_{i+1} = \left\{ \alpha \mid \alpha \in \Lambda_i \text{ and } |F(x_i, \alpha) - y^i| < \frac{\varepsilon}{2} \right\}.$$

Then the corridor at the $(i+1)$ th step will be $R_\varepsilon^{\Lambda_{i+1}}$.

The process is arranged to run so that, despite the decrease in the sets

$$A, \mathfrak{z} A, \mathfrak{z} \cdots \mathfrak{z} A,,$$

the quantities $C_\varepsilon^{\Lambda_i}$ preserve their values; that is,

$$C_\varepsilon^{\Lambda_i} = C_\varepsilon^{\Lambda_{i+1}}$$

and, in consequence, Λ_i is nonempty for all i 's. This follows from the fact that, by definition of $D_x^{\Lambda_i}$, the narrowing of Λ_i by the condition

$$|F(x, \alpha) - y| < \frac{\varepsilon}{2} \quad \text{for } y \in D_x^{\Lambda_i}$$

leaves $C_\varepsilon^{\Lambda_i}$ unchanged. Noting further that

$$D_x^{\Lambda_i} \supset D_x^{\Lambda_{i+1}}$$

and taking into account the result of Theorem 16.2, we get

$$\int \ln K_x^{\Lambda_i} dP_x \geq \int \ln K_x^{\Lambda_{i+1}} dP_x \geq C_\varepsilon^{\Lambda_i}. \quad (16.25)$$

We let the numerical set

$$D_x^0 = \bigcap_{i=1}^{\infty} D_x^{\Lambda_i}$$

correspond to each point $x \in X$. Accordingly,

$$K_x^0 = \lim_{\ell \rightarrow \infty} K_x^{\Lambda_\ell}.$$

It follows from Theorem 16.1 that D_x^0 is the union of a finite number of nonintersecting intervals, segments, or semi-intervals of a length not less than ε and belongs to the segment $[l - E, 1 + \varepsilon]$. Therefore,

$$K_x^0 \geq \varepsilon.$$

If the $K_x^{\Lambda_\ell}$ are measurable functions, then K_x^0 is measurable too.

Furthermore, since the $\ln K_x^{\Lambda_\ell}$ are uniformly modulo-bounded, it follows that for any realization of the process we have

$$\int K_x^0 dP_x = \int [\lim_{\ell \rightarrow \infty} \ln K_x^{\Lambda_\ell}] dP_x = \lim_{\ell \rightarrow \infty} \int \ln K_x^{\Lambda_\ell} dP_x.$$

Thus, any realization of the process A enables one to find D_x^0 satisfying requirement 1 of Theorem 16.3. We will show now that almost any realization of process A generates D_x^0 satisfying requirement 2 of Theorem 16.3.

Lemma 16.7. *Let*

$$x_1, \dots, x_k, x_{k+1}, \dots, x_{\ell+k}$$

be a sample of length $\ell + k$ from X , let the numbers y^1, \dots, y^k be sampled in the course of the process A, and let Λ^{k+1} be the subset A constructed by the $(k+1)$ th step of process. Consider the set R_ℓ^k specified by direct product

$$R_\ell^k = \prod_{i=k+1}^{k+\ell} D_{x_i}^{\Lambda_{k+1}},$$

and introduce a uniform density,

$$\mu_0(y) = \left[\prod_{i=k+1}^{k+\ell} K_{x_i}^{\Lambda_{k+1}} \right]^{-1},$$

on that set. Consider the subset $G_\ell^k \in R_\ell^k$ consisting of sequences $y^{k+1}, \dots, y^{k+\ell}$ such that $y^i \in D_{x_i}^{\Lambda_k}$ and the system of inequalities

$$|y^i - F(x_i, \alpha)| < \frac{\epsilon}{2}, \quad i = k+1, \dots, k+\ell \quad (16.26)$$

is resolvable for $\alpha \in \Lambda_k$.

Then for any fixed $\ell \geq 1$ the equality holds true

$$\lim_{k \rightarrow \infty} E \int_{G_\ell^k} \mu_0(y) dV = 1.$$

(Here the expected value is taken for all realizations of the process up to k th step and for the extensions of the sample $x_{k+1}, \dots, x_{k+\ell}$).

Proof. To begin with, we assume that

$$x_1, \dots, x_k, x_{k+1}, \dots, x_{k+\ell}$$

and that the sequence

$$y^1, \dots, y^k$$

is fixed. Then on any extension of the process A, as far as the $(k+\ell)$ th step, the sequences

$$y^1, \dots, y^{k+1}$$

will be sampled with a certain density μ_1 . We denote by T_ℓ^k the carrier of that density.

From the definition of the process A and of the density μ_0 , it follows that $T_\ell^k \subset R_\ell^k$ and that on T_ℓ^k the ratio of the densities μ_0/μ_1 is defined by

$$\frac{\mu_0}{\mu_1} = \prod_{i=k+1}^{k+\ell} \frac{K_{x_i}^{\Lambda_i}}{K_{x_i}^{\Lambda_{k+1}}} \quad (16.27)$$

Note that μ_1 does not depend on y for $y \in R_\ell$, but μ_0 does.

For

$$y^{k+1}, \dots, y^{k+\ell}$$

the system (16.26) is obviously a simultaneous one (the set of its solutions is just $\Lambda_{k+\ell+1}$). Therefore,

$$G_\ell^k \supset T_\ell^k.$$

Hence

$$\int_{G_\ell^k} \mu_0 dV \geq \int_{T_\ell^k} \mu_0 dV = \int_{T_\ell^k} \left(\frac{\mu_0}{\mu_1} \right) \mu_1 dV.$$

By taking the logarithm of the above inequality, we get

$$\ln \int_{G_\ell^k} \mu_0 dV \geq \ln \int_{T_\ell^k} \left(\frac{\mu_0}{\mu_1} \right) \mu_1 dV \geq \int_{T_\ell^k} \ln \left(\frac{\mu_0}{\mu_1} \right) \mu_1 dV.$$

Therefore using (16.27) we obtain

$$\begin{aligned} \ln \int_{G_\ell^k} P_0 dV &\geq \int_{T_\ell^k} \sum_{i=k+1}^{P+k} \left(\ln K_{x_i}^{\Lambda_i} - \ln K_{x_i}^{\Lambda_{k+1}} \right) \mu_1 dV \\ &= \sum_{i=k+1}^{\ell+k} \left(\int_{T_\ell^k} \ln K_{x_i}^{\Lambda_i} \mu_1 dV - \int \ln K_{x_i}^{\Lambda_{k+1}} dP_x \right). \end{aligned}$$

We now average the above inequality over all realizations of the process as far as the k th step and over all extensions $x_{k+1}, \dots, x_{k+\ell}$:

$$E \ln \int_{G_\ell^k} \mu_0 dV \geq \sum_{i=k+1}^{\ell+k} \left(E \int_{T_\ell^k} \ln K_{x_i}^{\Lambda_i} \mu_1 dV - E \int \ln K_{x_i}^{\Lambda_{k+1}} dP_x \right). \quad (16.28)$$

We denote

$$w_i = E \int \ln K_{x_i}^{\Lambda_i} dP_x,$$

where averaging is done over all realizations of the process as far as the i th step. By (16.25), we find that w_i is a decreasing sequence bounded from below. Therefore

$$\lim_{k \rightarrow \infty} |w_{\ell+k} - w_k| = 0.$$

Note also that for $k < i < k + \ell$ we have

$$E \int_{T_\ell^k} \ln K_{x_i}^{\Lambda_i} \mu_1 dV = w_i$$

In the above notation, the inequality (16.28) takes the form

$$E \ln \int_{G_\ell^k} \mu_0 dV \geq \sum_{i=k+1}^{k+\ell} (w_i - w_{k+1}) \geq -\ell |w_{k+1} - w_{\ell+k}| \rightarrow 0$$

for k tending to infinity and for a fixed ℓ .

Finally, using the inequality

$$x \geq 1 + \ln x,$$

we get

$$E \int_{G_\ell^k} \mu_0 dV \geq 1 + E \ln \int_{G_\ell^k} \mu_0 dV \rightarrow 1.$$

Lemma 16.7 has been proved.

Note. If we denote by \hat{G}_ℓ^k the complement to G_ℓ^k in R_ℓ^k —that is, the set of sequences

$$y^{k+1}, \dots, y^{k+\ell}, \quad y^i \in D_{x_k}^{\Lambda_k},$$

for which the system (16.26) is nonsimultaneous—then for k tending to infinity and for a fixed ℓ it is true that

$$\int_{\hat{G}_\ell^k} \mu_0 dV \xrightarrow{k \rightarrow \infty} 0.$$

Continued Proof of Theorem 16.3. Let

$$D_x^0 = \bigcap_{i=1}^{\ell} D_{x_i}^{\Lambda_i},$$

where $D_{x_i}^{\Lambda_i}$ are obtained in the course of the process. By the time the process is completed, let an additional sample

$$\bar{x}_1, \dots, \bar{x}_\ell$$

be collected. We denote by $T_\ell \in E_\ell$ the direct product

$$T_\ell = \prod_{i=1}^{\ell} D_{\bar{x}_i}^0$$

We introduce a uniform distribution on it specified by the density

$$\mu(y^1, \dots, y^\ell) = \begin{cases} 0 & \text{if } y \notin T_\ell, \\ \frac{1}{\prod_{i=1}^{\ell} K_{\bar{x}_i}^{\Lambda_i}} & \text{if } y \in T_\ell. \end{cases}$$

Let G_ℓ be the subset T_ℓ of sequence y^1, \dots, y^ℓ , such that the system

$$|F(\bar{x}_i, \alpha) - y^i| < \frac{\varepsilon}{2}, \quad i = 1, \dots, \ell \quad (16.29)$$

is nonsimultaneous for $\alpha \in A$.

Theorem 16.3 will have been proved if we establish that

$$E \int_{\tilde{G}_\ell^k} \mu dV = 0,$$

where averaging is done over all realizations of the process A and over all samples $\bar{x}_1, \dots, \bar{x}_\ell$.

We consider, as in Lemma 16.7, the uniform distribution μ_0 on the samples y^1, \dots, y^ℓ from the direct product

$$R_\ell^k = \prod_{i=1}^{\ell} D_{\bar{x}_i}^{\Lambda_{k+1}}. \quad (16.30)$$

Since $D_x^0 \subset D_{\bar{x}}^{\Lambda_k}$, it follows that $T_\ell \subset R_\ell^k$.

Now, as in Lemma 16.7, we denote by G_ℓ^k the subset R_ℓ^k for which the system (16.29) is simultaneous for $\alpha \in \Lambda_k$, and we denote by \tilde{G}_ℓ^k the complement to G_ℓ^k in R_ℓ^k . From (16.30) and by the definition of \tilde{G}_ℓ and \tilde{G}_ℓ^k it follows that

$$\tilde{G}_\ell^k \supset \tilde{G}_\ell.$$

Then

$$\begin{aligned} \int_{\tilde{G}_\ell} \mu dV &= \int_{\tilde{G}_\ell} \left(\frac{\mu}{\mu_0} \right) \mu_0 dV = \frac{\prod_{i=1}^{\ell} K_{\bar{x}_i}^{\Lambda_k}}{\prod_{i=1}^{\ell} K_{\bar{x}_i}^0} \int_{\tilde{G}_\ell^k} \mu_0 dV \\ &\leq \left(\frac{1+\varepsilon}{\varepsilon} \right)^{\ell} \int_{\tilde{G}_\ell^k} \mu_0 dV, \end{aligned}$$

because K_x^0 and $K_x^{\Lambda_k}$ are bounded from above and below by ε and $(1+\varepsilon)$.

By averaging the above inequality over all realizations of the process A and over all extensions $\bar{x}_1, \dots, \bar{x}_\ell$, we get

$$E \int_{\tilde{G}_\ell} \mu dV \leq \left(\frac{1+\varepsilon}{\varepsilon} \right)^{\ell} E \int_{\tilde{G}_\ell^k} \mu_0 dV.$$

On strength of the note to Lemma 16.7, the right-hand side of the inequality tends to zero with k tending to infinity and with ℓ fixed in advance; and since

this inequality holds true for any k , it follows that for all $\ell \geq 1$ we obtain

$$E \int_{\tilde{G}_\ell} \mu dV = 0.$$

This completes the proof of Theorem 16.3.

Corollary 1. *From conclusion 2 of Theorem 16.3 it follows that*

$$\int \ln K_x^0 dP_x = C_\varepsilon^\Lambda.$$

Proof. Indeed, from the definition of $V_\varepsilon(x_1, \dots, x_\ell)$ and from item 2 of Theorem 16.3 it follows that for almost all x_1, \dots, x_ℓ we obtain

$$V_\varepsilon(x_1, \dots, x_\ell) \geq \prod_{i=1}^{\ell} K_{x_i}^0.$$

Hence

$$\frac{E \ln V_\varepsilon^\Lambda(x_1, \dots, x_\ell)}{\ell} \geq E \ln K_{x_i}^0.$$

On passing to the limit, we obtain

$$C_\varepsilon^\Lambda = \lim_{\ell \rightarrow \infty} \frac{E \ln V_\varepsilon^\Lambda(x_1, \dots, x_\ell)}{\ell} \geq \int \ln K_x^0 dP_x.$$

Taken in conjunction with the statement of item 1 in Theorem 16.3, this proves Corollary 1.

Corollary 2. *For the class of functions $F(x, a)$, $a \in A$, and for the measure P_x , let there hold true the inequality*

$$C_\varepsilon^\Lambda = \ln \varepsilon + \eta, \quad \eta > 0.$$

Suppose here that there is specified another measure, P_x^ , absolutely continuous with respect to P_x and let the density*

$$p(x) = \frac{dP_x^*}{dP_x}$$

satisfy the condition

$$p(x) > a > 0.$$

Then

$$\bar{C}_\varepsilon^\Lambda \geq \ln \varepsilon + a\eta,$$

where

$$\bar{C}_\varepsilon^\Lambda = \lim_{\ell \rightarrow \infty} E_{P_x^*} \ln V(x_1, \dots, x_\ell).$$

Proof. Indeed, the inequality

$$V_\varepsilon^\Lambda(x_1, \dots, x_\ell) \geq \prod_{i=1}^\ell K_{x_i}^0$$

is satisfied for almost all x_1, \dots, x_ℓ in the case of the measure P_x^* as well. By reasoning along the same lines as in Corollary 1, we obtain

$$\frac{E_{P_x^*} \ln V_\varepsilon^\Lambda(x_1, \dots, x_\ell)}{\ell} \geq \int \ln K_{x_i}^0 dP_x^*.$$

Furthermore, recalling that

$$\ln K_x^0 - \ln \varepsilon \geq 0,$$

we get

$$\begin{aligned} \int \ln K_x^0 dP_x^* &= \int \ln \varepsilon dP_x^* + \int [\ln K_x^0 - \ln \varepsilon] dP_x^* \\ &= \ln \varepsilon + \int [\ln K_x^0 - \ln \varepsilon] p(x) dP_x \\ &\geq \ln \varepsilon + a[\bar{C}_\varepsilon^\Lambda - \ln \varepsilon] = \ln \varepsilon + a\eta. \end{aligned}$$

Corollary 3. *Let*

$$C_\varepsilon^\Lambda > \ln \varepsilon.$$

Also let P_x and P_x^ be absolutely continuous with respect to each other. Then*

$$\bar{C}_\varepsilon^\Lambda > \ln \varepsilon.$$

Proof. Let

$$C_\varepsilon^\Lambda = \int \ln K_x^0 dP_x > \ln \varepsilon$$

hold true. Denote

$$I = \{x : \ln K_x^0 > \ln \varepsilon\}.$$

Then $P(I) > 0$ and, in consequence, $P^*(I) > 0$. Therefore,

$$\bar{C}_\varepsilon^\Lambda \geq \int \ln K_x^0 dP_x^* \geq \ln \varepsilon + \int_I [\ln K_x^0 - \ln \varepsilon] dP_x^* \geq \ln \varepsilon.$$

Corollary 4. *For almost all realizations of the process A, for all $4 > 1$, for almost all samples x_1, \dots, x_ℓ , for all $y^i \in \bar{D}_{x_i}^0$ (where $\bar{D}_{x_i}^0$ is the closure $D_{x_i}^0 = \bigcap_{k=1}^{\infty} D_{x_i}^{\Lambda_k}$), for all $\delta > 0$, and for $k \geq 1$, one can find $\alpha^* \in \Lambda_k$ such that for all i's ($1 \leq i \leq \ell$) we obtain*

$$|F(x_i, \alpha^*) - y^i| \leq \frac{\varepsilon}{2} + \delta$$

with

$$\int \ln K_x^0 dP_x = C_\varepsilon^\Lambda.$$

16.5 THEOREM ON THE EXISTENCE OF FUNCTIONS CLOSE TO THE CORRIDOR BOUNDARIES (THEOREM ON POTENTIAL NONFALSIFIABILITY)

Theorem 16.4. *Let*

$$C_\varepsilon^\Lambda = \ln \varepsilon + \eta, \quad \eta > 0.$$

Then there exist functions

$$\psi_1(x) \geq \psi_0(x)$$

that possess the following properties:

1. $\int |\psi_1(x) - \psi_0(x)| dP_x \geq \varepsilon(e^\eta - 1).$
2. *For any given $\delta > 0$, one can show a subclass $A^* \subset A$ such that $C_\varepsilon^{A^*} = C_\varepsilon^\Lambda$, and one can also show functions*

$$\phi_1 = \sup_{\alpha \in \Lambda^*} Q(x, \alpha), \quad \phi_0 = \inf_{\alpha \in \Lambda^*} Q(x, \alpha)$$

such that

$$\phi_0(x) \leq \psi_0(x), \quad \phi_1(x) \geq \psi_1(x).$$

$$3. \quad \int (\phi_1(x) - \psi_1(x)) dP_x < \delta,$$

$$\int (\psi_0(x) - \phi_0(x)) dP_x < \delta.$$

4. For any $\delta_1 > 0$ and $\ell \geq 1$, for almost any sequence x_1, \dots, x_ℓ , and for any sequence $\omega_1, \dots, \omega_\ell$ ($w = 0, 1$), one can find $\alpha^* \in A^*$ such that

$$|Q(x_i, \alpha^*) - \psi_{\omega_i}(x_i)| < \delta_1, \quad i = 1, \dots, \ell.$$

Proof. We will show that almost any realization of the process A described in the previous section allows one to find the required functions $\psi_1(x)$ and $\psi_0(x)$ —if one sets

$$\psi_1(x) = \sup_{y \in D_x^0} y - \frac{\varepsilon}{2}, \quad \psi_0(x) = \inf_{y \in D_x^0} y + \frac{\varepsilon}{2} \quad (16.31)$$

and uses as A^* the subclass A^k generated at the k th step of the process, where k is chosen according to δ and the specific realization of the process.

To prove this we need some intermediate results.

Lemma 16.8. Let Y and Y^* ($Y^* \subset Y$) be two open sets in Euclidean space E_k . Let their volumes, V and V^* , respectively, be finite and $V^* < V$. Suppose that the density $p(x)$, whose carrier T belongs to Y , has a finite entropy

$$H = - \int p(x) \ln p(x) dV.$$

Then the estimate

$$p^* < \frac{\ln V - H + \ln 2}{\ln V - \ln V^*},$$

where

$$p^* = \int_Y p(x) dV,$$

holds true.

Proof of Lemma 16.8. The entropy H will be a maximum if $P(Y^*) = p^*$ and in the case of a uniform distribution on Y^* and Y/Y^* —that is, when

$$p(x) = \begin{cases} \frac{p^*}{V} & \text{if } x \in Y^*, \\ \frac{(1-p^*)}{V-V^*} & \text{if } x \in Y/Y^*. \end{cases}$$

Therefore

$$\begin{aligned} H &\leq -p^* \ln \frac{p^*}{V^*} - (1-p^*) \ln \frac{(1-p^*)}{V-V^*} \\ &= -p^* \ln p^* - (1-p^*) \ln (1-p^*) + p^* \ln V^* + (1-p^*) \ln (V - V^*) \\ &\leq \ln 2 + \ln V - p^* (\ln V - \ln V^*). \end{aligned}$$

Hence

$$p^* \leq \frac{\ln V - H + \ln 2}{\ln V - \ln V^*}.$$

Lemma 16.8a. Let for some point x and some number b the inequality

$$C_{\varepsilon,x}^{\Lambda} < C_{\varepsilon}^{\Lambda}$$

holds true. We define the set

$$B_k(x, b) = \left\{ \alpha : \alpha \in \Lambda_k \text{ and } b - \frac{\varepsilon}{2} \leq F(x, \alpha) \leq b + \frac{\varepsilon}{2} \right\},$$

where Λ_k is the subclass constructed by the k th step of the process A. Then

$$P\{B_k(x, b) \neq \emptyset\} \xrightarrow{k \rightarrow \infty} 0$$

(the probability is taken over the set of realizations of the process A).

In other words, the probability that the system of inequality

$$|F(x, \alpha) - b| < \frac{\varepsilon}{2}, \quad |F(x_i, \alpha) - y^i| < \frac{\varepsilon}{2}, \quad i = 1, \dots, k, \quad (16.32)$$

for $\alpha \in \Lambda$ will remain simultaneous tends to zero in the course of the process A.

Proof: We set $t < k$, and we fix the sample $x_1, \dots, x_t, x_{t+1}, \dots, x_k$ and the sequence y^1, \dots, y^t obtained in the course of the process A. We consider, as in Lemma 16.7, the density μ_1 on the sequence y^{t+1}, \dots, y^k in

$$R_k^t = \prod_{i=t+1}^k D_{x_i}^{\Lambda_{i+1}},$$

generated by the process A on steps $t+1, \dots, k$.

We further denote by $Y_{x,\varepsilon}^{\Lambda_t}(b, x_{t+1}, \dots, x_k)$ the subset R_k^t consisting of sequences y^{t+1}, \dots, y^k such that the system (16.32) is simultaneous for $\alpha \in \Lambda$. Then

$$P\{B_k(x, b) \neq \emptyset\} = E \left\{ \int_{x_{t+1}, \dots, x_k} \left[\int_{Y_{x,\varepsilon}^{\Lambda_t}} \mu_1 dV \right] dP_{x_{t+1}, \dots, x_k} \right\}, \quad (16.33)$$

where averaging E is taken over the process as far as the t th step.

We denote by V the volume of the set R_k^t , and we denote by V^* the volume $Y_{x,\epsilon}^{\Lambda_t}(b, x_{t+1}, \dots, t_k)$. Then

$$\frac{1}{k} \ln V \xrightarrow[k \rightarrow \infty]{P(x_{t+1}, \dots, x_k)} E_x \ln K_x^{\Lambda_t} \geq C_\epsilon^\Lambda$$

and, owing to the condition of Lemma 16.8a, we obtain

$$\frac{1}{k} \ln V^* \xrightarrow[k \rightarrow \infty]{P(x_{t+1}, \dots, x_k)} C_{x,\epsilon}^{\Lambda_t}(b) < C_\epsilon^\Lambda.$$

Therefore, on setting

$$C_1 = \frac{C_\epsilon^\Lambda - C_{x,\epsilon}^{\Lambda_t}(b)}{2} > 0,$$

we get

$$P_{x_{t+1}, \dots, x_k} \left\{ \left(\frac{1}{k} \ln V - \frac{1}{k} \ln V^* \right) < C_1 \right\} \xrightarrow{k \rightarrow \infty} 0. \quad (16.34)$$

We now estimate

$$I_t^k = \int_{x_{t+1}, \dots, x_k} \left[\int_{Y_{x,\epsilon}^{\Lambda_t}} \mu_1 dV \right] dP_x,$$

recalling that, with t specified in advance, V and V^* depend solely on x_{t+1}, \dots, x_k :

$$\begin{aligned} I_t^k &= \int_{(1/k)(\ln V - \ln V^*) \geq C_1} \left[\int_{Y_{x,\epsilon}^{\Lambda_t}} \mu_1 dV \right] dP_x \\ &\quad + \int_{(1/k)(\ln V - \ln V^*) < C_1} \left[\int_{Y_{x,\epsilon}^{\Lambda_t}} \mu_1 dV \right] dP_x. \end{aligned}$$

We denote by I_1 and I_2 the terms on the right-hand side. Then

$$I_2 \leq P_{x_{t+1}, \dots, x_k} \left\{ \frac{1}{k} (\ln V - \ln V^*) < C_1 \right\}.$$

Also, by virtue of Lemma 16.8 and on replacing $Y = R_t$ and $Y^* = Y_{x,\epsilon}^{\Lambda_t}$, $p(x) = \mu_1$, we obtain

$$\begin{aligned} I_1 &\leq \int_{(1/k)(\ln V - \ln V^*) \geq C_1} \frac{(1/k)(\ln V - H + \ln 2)}{(1/k)(\ln V - \ln V^*)} dP_x \\ &\leq \int_{x_{t+1}, \dots, x_k} \frac{(1/k)(\ln V - H + \ln 2)}{C_1} dP_x. \end{aligned}$$

Here H is the entropy of the distribution μ_1 on the sequence y^{t+1}, \dots, y^k , with x_{t+1}, \dots, x_k specified in advance.

Note that $H \leq \ln V$, so the extension of the integral to the entire space x_{t+1}, \dots, x_k is quite legitimate. Furthermore, since

$$\lim_{k \rightarrow \infty} I_2 = 0$$

by virtue of (16.34) and since

$$E_{x_{t+1}, \dots, x_k} \frac{\ln V}{k} = \int K_x^{\Lambda_t} dP_x$$

by definition, in view of Corollary 4 of Theorem 16.3 we obtain

$$\lim_{k \rightarrow \infty} \frac{1}{\ell} E_{x_{t+1}, \dots, x_k} H = C_\varepsilon^{\Lambda_t}.$$

Therefore, for fixed t and A , we have

$$\lim_{k \rightarrow \infty} I_t^k \leq \frac{1}{C_1} \left(\int K_x^{\Lambda_t} dP_x - C_\varepsilon^{\Lambda_t} \right).$$

On inserting the above result in (16.33) and noting that the function under the integral is bounded, we obtain

$$\lim_{k \rightarrow \infty} P\{B_k(x, b) \neq \emptyset\} = E(\lim_{k \rightarrow \infty} I_t^k) \leq \frac{1}{C_1} \left(E \int K_x^{\Lambda_t} dP_x - C_\varepsilon^{\Lambda_t} \right),$$

where E is taken over the realizations of the process A . This estimate is valid for any $t \geq 1$ and in view of Corollary 4 of Theorem 16.3:

$$\lim_{t \rightarrow \infty} E \int K_x^{\Lambda_t} dP_x = C_\varepsilon^{\Lambda_t}.$$

Therefore

$$\lim_{k \rightarrow \infty} P\{B_k(x, b) \neq \emptyset\} = 0.$$

Corollary 1. Suppose that for some $x \in X$ and some number b , the inequality

$$C_{x, \varepsilon}^{\Lambda}(y) < C_\varepsilon^{\Lambda}$$

hold true for all $y \geq b$. Then for

$$\bar{B}_k(x, b) = \{a : a \in \Lambda \text{ and } F(x, a) > b - \frac{\varepsilon}{2}\}$$

the following relation holds true:

$$\lim_{k \rightarrow \infty} P\{\bar{B}_k(x, b) \neq \emptyset\} = 0.$$

Proof. By taking a decreasing finite sequence of numbers b_1, \dots, b_s such that

$$b_1 = 1 + \varepsilon, \quad b_s = b, \quad b_i - b_{i+1} < \varepsilon,$$

we get

$$\bar{B}_k(x, b) = \bigcup_{i=1}^s B_k(x, b_i),$$

where $B_k(x, h)$ is defined in the conditions of Lemma 16.8 and all numbers b_i satisfy the conditions of Lemma 16.8. Therefore

$$P\{\bar{B}_k(x, b) \neq \emptyset\} \leq \sum_i^s P\{B_k(x, b_i) \neq \emptyset\} \xrightarrow{k \rightarrow \infty} 0.$$

We have thus proved Corollary 1.

Corollary 2. If, for some \mathbf{x} , a number h , and the step k_0 of the process A ,

$$C_{x,\varepsilon}^{\Lambda_{k_0}}(y) < C_{x,\varepsilon}^{\Lambda}(y) < C_x^{\Lambda}$$

for all $y > 0$, then for

$$\bar{B}_k^{\Lambda_{k_0}}(x, b) = \left\{ \mathbf{a} : \mathbf{a} \in \Lambda_{k_0} \text{ and } F(x, \mathbf{a}) > b - \frac{\varepsilon}{2} \right\},$$

it is true that

$$P\{\bar{B}_k^{\Lambda_{k_0}}(x, b) \neq \emptyset\} \xrightarrow{k \rightarrow \infty} 0,$$

where the probability is determined over all extension of the process beyond step k_0 .

Lemma 16.9. Let $\psi_1(\mathbf{x})$ and $\psi_0(\mathbf{x})$ be defined, according to (16.31). Consider

$$\phi_1^k(x) = \sup_{\alpha \in \Lambda_k} F(x, \alpha), \quad \phi_0^k(x) = \inf_{\alpha \in \Lambda_k} F(x, \alpha).$$

Then for any \mathbf{x} and almost any realization of the process A we obtain

$$\phi_1^k(x) \xrightarrow{k \rightarrow \infty} \psi_1(x), \quad \text{and} \quad \phi_0^k(x) \xrightarrow{k \rightarrow \infty} \psi_0(x),$$

with

$$\phi_1^k(x) \geq \psi_1(x) \quad \text{and} \quad \psi_0(x) \geq \phi_0^k(x).$$

Proof. We will prove Lemma 16.9 for $\phi_1(x)$ and $\psi_1(x)$. The case $\phi_0(x)$ and $\psi_0(x)$ can be proved in a similar way. We denote

$$\bar{\psi}_\varepsilon^k(x) = \sup_{y \in D_x^{\Lambda_k}} y.$$

Then, by definition,

$$\psi_1(x) = \lim_{k \rightarrow \infty} \bar{\psi}_\varepsilon^k(x) - \frac{\varepsilon}{2}.$$

Corollary 2 of Lemma 16.8 may be stated thus:

If, for some x , b , and k , the condition

$$\bar{\psi}_\varepsilon^{k_0}(x) < b$$

is satisfied, then

$$P \left\{ \phi_1^k(x) > b - \frac{\varepsilon}{2} \right\}_{k \rightarrow \infty} \rightarrow 0.$$

On the other hand,

$$\phi_1^k(x) > \bar{\psi}_\varepsilon^{k_0}(x) - \frac{\varepsilon}{2}$$

because if $y \in D_x^{\Lambda_k}$, then there exists $\alpha \in \Lambda_k$ such that

$$F(x, \alpha) > y - \frac{\varepsilon}{2}.$$

Therefore we get

$$\phi_1^k(x) \xrightarrow[k \rightarrow \infty]{P} \psi(x).$$

Furthermore, since $\phi_1^k(x)$ is a monotone decreasing sequence bounded from below, it follows that the convergence almost surely stems from convergence in probability.

Corollary 1. *For almost any realization of the process A, it is true that*

$$\phi_1^k(x) \rightarrow \psi_1(x) \quad \text{and} \quad \phi_0^k(x) \rightarrow \psi_0(x)$$

almost everywhere in X, as k tends to infinity.

Corollary 2. For almost any realization of the process A, the functions $\phi_0^k(x)$ and $\phi_1^k(x)$ converge in the mean with respect to X towards $\psi_0(x)$ and $\psi_1(x)$, respectively; that is,

$$\lim_{k \rightarrow \infty} \int |\phi_0^k(x) - \psi_0(x)| dP_x = 0,$$

$$\lim_{k \rightarrow \infty} \int |\phi_1^k(x) - \psi_1(x)| dP_x = 0.$$

Proof. This result stems immediately from Corollary 1 in view of the fact that integrands are bounded.

Continued Proof of Theorem 16.4. To complete the proof, it remains to combine the results of Corollary 4 of Theorem 16.3 and Corollaries 1 and 2 of Lemma 16.9.

Indeed, the conclusion of those corollaries holds true for any realization of the process A simultaneously. By choosing one such realization, we get the following:

$$1. \quad \psi_1(x) \geq \psi_0(x),$$

$$\int [\psi_1(x) - \psi_0(x)] dP_x = \int |K_x^0 - \varepsilon| dP_x = \int K_x^0 dP_x - \varepsilon,$$

because $K_x^0 \geq \varepsilon$.

But (see Corollary 4 of Theorem 16.3)

$$\ln \int K_x^0 dP_x \geq \int \ln K_x^0 dP_x = C_\varepsilon^\Lambda.$$

Hence.

$$\int |\psi_1(x) - \psi_0(x)| dP_x \geq \exp \{C_\varepsilon^\Lambda\} - \varepsilon = \varepsilon(e^\eta - 1).$$

Requirement 1 is thus satisfied.

2. By virtue of Corollary 2 of Lemma 16.9, one can, proceeding from the specified $\delta > 0$, choose k such that

$$\int |\phi_1^k(x) - \psi_1(x)| dP_x < \delta,$$

$$\int |\phi_0^k(x) - \psi_0(x)| dP_x < \delta$$

with $\phi_1^k(x) \geq \psi_1(x)$ and $\phi_0^k(x) \leq \psi_0(x)$. By using Λ_k as A^* we satisfy requirements 2 and 3.

3. By Corollary 2 of Lemma 16.9, for almost any sequence x_1, \dots, x_ℓ , one can find for the specified $\delta_1 > 0$ the value $k_1 > k$ such that

$$\begin{aligned} |\phi_1^{k_1}(x_i) - \psi_1(x_i)| &< \delta_1, \\ |\phi_0^{k_1}(x_i) - \psi_0(x_i)| &< \delta_1 \end{aligned} \quad (16.35)$$

simultaneously for all i's ($1 \leq i \leq \ell$).

Now let

$$\omega_1, \dots, \omega_\ell$$

be any sequence of 0's and 1's. We set

$$y_i = \begin{cases} \psi_{\omega_i}(x_i) + \frac{\varepsilon}{2} & \text{if } \omega_i = 1, \\ \psi_{\omega_i}(x_i) - \frac{\varepsilon}{2} & \text{if } \omega_i = 0. \end{cases}$$

Then $y_i \in \bar{D}_x^0$, where \bar{D}_x^0 is the closure of D_x^0 . By Corollary 4 of Theorem 16.3, one can find

$$\alpha^* \subset \Lambda_k \subset \Lambda_{k_1} \subset \Lambda^*$$

such that

$$|F(x_i, \alpha^*) - y^i| < \frac{\varepsilon}{2} + \delta_1 \quad (16.36)$$

for all i's ($1 \leq i \leq \ell$).

Furthermore, let $\omega_i = 0$ for, say, some i. Then

$$F(x_i, \alpha^*) \geq \inf_{\alpha \in \Lambda_{k_1}} F(x_i, \alpha) = \phi_0^{k_1} > \psi_0(x_i) - \delta_1$$

and

$$F(x_i, \alpha^*) < y^i + \frac{\varepsilon}{2} + \delta_1 = \psi_0(x_i) + \delta_1;$$

that is,

$$|F(x_i, \alpha^*) - \psi_{\omega_i}(x_i)| < \delta_1.$$

The case $\omega_i = 1$ is proved similarly. Requirement 4 is satisfied.

We have thus proved Theorem 16.4.

16.6 THE NECESSARY CONDITIONS

Theorem 165. *For one-sided uniform convergence to take place it is necessary that, for any $\varepsilon > 0$, there should exist a finite ε net in the metric $L_1(P)$ of the set of functions $F(x, \alpha)$, $\alpha \in A$. (That is a finite collection $\alpha_1, \dots, \alpha_N$ such that for any $\alpha^* \in A$ one can find α_k for which the inequality*

$$\int |F(x, \alpha^*) - F(x, \alpha_k)| dP_x \leq \varepsilon \quad (16.37)$$

is satisfied.)

For (nontrivial) consistency of the maximum likelihood method, an analogous theorem is valid .

Theorem 16.5a. *Let $p(x, \alpha)$, $\alpha \in A$ be a set of densities, satisfying the condition*

$$|\ln p(x, \alpha)| \leq B, \quad \alpha \in A.$$

For the maximum likelihood method to be consistent it is necessary that for any ε , there should exist a finite collection $\alpha_1, \dots, \alpha_N$, such that for any $\alpha^ \in A$ the following inequality*

$$\int |\ln p(x, \alpha^*) - \ln p(x, \alpha_k)| dP_x \leq \varepsilon$$

is satisfied for at least one k ($1 \leq k \leq N$).

Lemma 16.10. *For the class of functions $F(x, \alpha)$, $\alpha \in A$, let one-sided uniform convergence take place, that is,*

$$\sup_{\alpha \in A} \left(EF(x, \alpha) - \frac{1}{\ell} \sum_{i=1}^{\ell} F(x_i, \alpha) \right) \xrightarrow[\ell \rightarrow \infty]{P} 0, \quad (16.38)$$

and let there exist a bounded function $\psi_0(x)$ such that for any $\delta_1 > 0$ and for almost any sequence x_1, \dots, x_ℓ there is $\alpha^ \in A$ such that for all i 's ($1 \leq i \leq \ell$) the inequalities hold*

$$|F(x_i, \alpha^*) - \psi_0(x_i)| < \delta_1. \quad (16.39)$$

Then

$$\inf_{\alpha \in A} EF(x, \alpha) \leq E\psi_0(x).$$

Proof. We choose $\delta_1 > 0$. Let x_1, \dots, x_ℓ be a sample. We set out to find a^* satisfying (16.39). Then

$$\begin{aligned} \sup_{\alpha \in \Lambda} \left(EF(x, \alpha) - \frac{1}{\ell} \sum_{i=1}^{\ell} F(x_i, \alpha) \right) &\geq EF(x, a^*) - \frac{1}{\ell} \sum_{i=1}^{\ell} F(x_i, a^*) \\ &= (EF(x, a^*) - E\psi_0(x)) + \left(E\psi_0(x) - \frac{1}{\ell} \sum_{i=1}^{\ell} \psi_0(x_i) \right) \\ &\quad + \frac{1}{\ell} \sum_{i=1}^{\ell} (\psi_0(x_i) - F(x_i, a^*)) \\ &\geq \left(\inf_{\alpha \in \Lambda} EF(x, \alpha^*) - E\psi_0(x) \right) + \left(E\psi_0(x) - \frac{1}{\ell} \sum_{i=1}^{\ell} \psi_0(x_i) \right) - \delta_1. \end{aligned}$$

By passing to the limit in probability, we get by (16.38) and by the law of large numbers for $\psi_0(x)$

$$0 \geq \inf_{\alpha \in \Lambda} EF(x, \alpha) - E\psi_0(x) - \delta_1,$$

that is,

$$\inf_{\alpha \in \Lambda} EF(x, \alpha) \leq E\psi_0(x) + \delta_1.$$

In view of the arbitrary choice of δ_1 , we have thus obtained the statement of Lemma 16.10.

For the maximum likelihood case the analogous lemma is valid .

Lemma 16.10a. *Let the maximum likelihood method be (nontrivial)consistent for a class of uniformly bounded densities $p(x, \alpha), \alpha \in A$, uniformly separable from zero; that is, for any α_0 we have*

$$\inf_{\alpha \in \Lambda} \frac{1}{\ell} \sum_{i=1}^{\ell} -\ln p(x_i, \alpha) \xrightarrow[\ell \rightarrow \infty]{P_{\alpha_0}} E_{\alpha_0}(-\ln p(x, \alpha_0)). \quad (16.38a)$$

Suppose also that there is a bounded function $\psi_0(x)$ such that for any $\delta_1 > 0$ and for almost any sequence x_1, \dots, x_ℓ (in the sense of the basic measure) there exists $a^* \in A$ such that for all i 's ($1 \leq i \leq \ell$) we have

$$|\ln p(x_i, a^*) + \psi_0(x_i)| \leq \delta_1. \quad (16.39a)$$

Then

$$E_{\alpha_0}(-\ln p(x, \alpha_0)) \leq E_{\alpha_0} \psi_0(x).$$

Proof. We choose $\delta_1 > 0$. Let x_1, \dots, x_ℓ be the sample and let $\alpha^* \in A$ satisfy (16.39a). Then

$$\begin{aligned} E_{\alpha_0}(-\ln p(x, \alpha_0)) &- \inf_{\alpha \in A} \frac{1}{\ell} \sum_{i=1}^{\ell} (-\ln p(x_i, \alpha)) \\ &\geq E_{\alpha_0}(-\ln p(x, \alpha_0)) - \frac{1}{\ell} \sum_{i=1}^{\ell} (-\ln p(x_i, \alpha^*)) \\ &= E_{\alpha_0}(-\ln p(x, \alpha_0)) - E_{\alpha_0} \psi_0(x) \\ &\quad + \left(E_{\alpha_0} \psi_0(x) - \frac{1}{\ell} \sum_{i=1}^{\ell} \psi_0(x_i) \right) + \frac{1}{\ell} \sum_{i=1}^{\ell} (\psi_0(x_i) + \ln p(x_i, \alpha^*)) \\ &\geq (E_{\alpha_0}(-\ln p(x, \alpha_0)) - E_{\alpha_0} \psi_0(x)) + \left(E_{\alpha_0} \psi_0(x) - \frac{1}{\ell} \sum_{i=1}^{\ell} \psi_0(x_i) \right) + \delta_1. \end{aligned}$$

By passing to the limit in probability and noting that (16.39a) is satisfiable for almost any sequence in the sense of the measure P_{α_0} , we obtain

$$0 \geq E_{\alpha_0}(-\ln p(x, \alpha_0)) - E_{\alpha_0} \psi_0(x) - \delta_1.$$

In view of the arbitrary choice of δ_1 , we have thus obtained the statement of Lemma 16.10a.

Auxiliary Statement. Let $F(x, \alpha)$, $\alpha \in A$, be a class of functions, let

$$\phi_0(x) = \inf_{\alpha \in A} F(x, \alpha)$$

be a measurable function, and let there exist a function $\psi_0(x)$ such that:

$$(a) \quad \phi_0(x) \leq \psi_0(x),$$

$$(b) \quad \int (\psi_0(x) - \phi_0(x)) dP_x < \delta,$$

$$(c) \quad \inf_{\alpha \in A} \int F(x, \alpha) dP_x < \int \psi_0(x) dP_x.$$

Then for any α the inequalities

$$(1) \quad \int |F(x, \alpha) - \phi_0(x)| dP_x \leq EF(x, \alpha) + E\psi_0(x) + 2\delta,$$

$$(2) \quad \inf_{\alpha \in A} \int |F(x, \alpha) - \psi_0(x)| dP_x \leq 2\delta$$

are valid.

Proof. Let $\mathbf{A} \in \mathbf{A}$. We denote

$$I = \{x : F(x, \alpha) < \psi_0(x)\}$$

Then

$$\begin{aligned} & \int |F(x, \alpha) - \psi_0(x)| dP_x \\ &= \int (F(x, \alpha) - \psi_0(x)) dP_x + 2 \int_I (\psi_0(x) - F(x, \alpha)) dP_x \\ &\leq \int (F(x, \alpha) - \psi_0(x)) dP_x + 2 \int (\psi_0(x) - \phi_0(x)) dP_x \\ &\leq \int (F(x, \alpha) - \psi_0(x)) dP_x + 2\delta. \end{aligned}$$

By applying the operator \inf to both sides of the inequality, we get

$$\inf_{\alpha \in \Lambda} \int |F(x, \alpha) - \psi_0(x)| dP_x \leq \inf_{\alpha \in \Lambda} EF(x, \alpha) - E\psi_0(x) + 2\delta.$$

Recalling point (c) above, we thus obtain the desired statement.

Proof of Theorem 16.5. Assume the reverse; that is, suppose that one-sided uniform convergence occurs, but there exists $\varepsilon_0 > 0$ for which there is no finite ε_0 net in $L_1(P)$.

Step 1. Decompose the class \mathbf{A} into finite number of subclasses $\Lambda_1, \dots, \Lambda_N$ such that for each of them the condition

$$\sup_{\alpha \in \Lambda_i} EF(x, \alpha) - \inf_{\alpha \in \Lambda_i} EF(x, \alpha) < \frac{\varepsilon_0}{3} \quad (16.40)$$

is satisfied. Obviously, this can be always done. Moreover, for at least one of those subclasses there will exist no network in $L_1(P)$. We denote it as A^* .

Step 2. In any set having no finite ε_0 net, it is possible to choose an infinite (countable) ε_0 -discrete subset—that is, such that for any two of its elements, x and y , we obtain

$$\rho(x, y) \geq \varepsilon_0.$$

We choose this ε_0 -discrete subset in A^* for the metric $L_1(P)$ and denote it as A^{**} . Obviously, this Λ^{**} will also have no finite ε_0 net.

Step 3. It is shown in Chapter 15 that the existence of a finite ε_0 net in $L_1(P)$ for any given ε_0 is a necessary condition for means to converge uniformly

to their expected values. Hence from Corollary 2 of Theorem 16.3 we conclude that for any A^{**} there exist positive numbers ε and η such that

$$C_\varepsilon^\Lambda > \ln \varepsilon + \eta.$$

Step 4. By Theorem 16.4, there exists a function $\psi_0(x)$ such that for $\delta = \varepsilon_0/2$ one can find a subclass $\Lambda_*^{**} \subset A^{**}$ for which

- (a) $C_\varepsilon^{\Lambda_*^{**}} = C_\varepsilon^\Lambda > \ln \varepsilon + \eta$,
- (b) $\psi_0(x) \geq \inf_{\alpha \in \Lambda_*^{**}} F(x, \alpha) = \phi_0(x)$,
- (c) $\int (\psi_0(x) - \phi_0(x)) dP_x \leq \delta$,
- (d) For any $\delta_1 > 0$ and for almost any sample x_1, \dots, x_ℓ there exists $a^* \in \Lambda_*^{**}$ such that

$$|F(x_i, a^*) - \psi_0(x_i)| < \delta_1.$$

Note that A^* inhibits the ε_0 -discrete properties in $L_1(P)$, the condition (16.40), and one-sided uniform convergence.

Step 5. On applying Lemma 16.10 to the class Λ_*^{**} (by virtue of Property 2), we obtain

$$\inf_{\alpha \in \Lambda_*^{**}} E F(x, \alpha) \leq E \psi_0(x).$$

Step 6. On applying the auxiliary statement (by virtue of (b) and (c) and Step 5), we obtain the following:

1. For any $a \in A^*$

$$E |F(x, a) - \psi_0(x)| < E(F(x, a) - \psi_0(x)) + 2\delta.$$

$$2. \quad \inf_{a \in \Lambda_*^{**}} E |F(x, a) - \psi_0(x)| \leq 2\delta.$$

Step 7. From statement 1 above we have

$$\inf_{\alpha \in \Lambda_*^{**}} |E F(x, \alpha) - E \psi_0(x)| \leq 2\delta.$$

Hence from (16.40) for any $a \in \Lambda_*^{**}$ we obtain

$$|E F(x, a) - |E \psi_0(x)|| < \frac{\varepsilon_0}{3} + 2\delta.$$

Inserting the $\delta = \varepsilon_0/4$, we obtain

$$E |F(x, a) - \psi_0(x)| \leq \frac{2\varepsilon_0}{3}.$$

Since the class Λ_*^{**} is ε_0 -discrete, we find that it consists of just one element. But this contradicts property (a) of Step 4 because for a one-element A it is always the case that

$$C_\varepsilon^A = \ln e.$$

This completes the proof of Theorem 16.5.

Proof of the Theorem 16.5a. This proof runs along approximately the same lines as that of Theorem 16.5. Let us trace it step by step. We denote $F(x, a) = -\ln p(x, a)$.

Step 1. Not mandatory. We assume $A^* = A$.

Step 2. Same as for Theorem 16.5.

Step 3. Same as for Theorem 16.5.

Step 4. Same as for Theorem 16.5 except that $\delta = \varepsilon_0 a / 8A$, where

$$\begin{aligned} a &= \inf_{\alpha, x} p(x, \alpha), & a > 0, \\ A &= \sup_{\alpha, x} p(x, a), & A < \infty. \end{aligned} \tag{16.41}$$

Step 5. By applying Lemma 16.10a to subclass Λ_*^{**} and recalling that the strict consistency of the maximum likelihood method is inherited for any subclass, we obtain for any $\alpha_0 \in \Lambda_*^{**}$

$$E_{\alpha_0} F(x, \alpha_0) \leq E_{\alpha_0} \psi_0(x).$$

Step 6. In view of property (c), Step 4, and (16.41), we have for $a \in \Lambda_*^{**}$

$$\int (\psi_0(x) - \phi_0(x)) p(x, \alpha_0) dP_x \leq A \int (\psi_0(x) - \phi_0(x)) dP_x < A\delta,$$

where

$$\phi_0(x) = \inf_{\alpha \in \Lambda_*^{**}} F(x, \alpha).$$

By applying the auxiliary statement for the measure P_{α_0} , we get

$$E_{\alpha_0} |F(x, \alpha_0) - \psi_0(x)| \leq E_{\alpha_0} F(x, \alpha_0) - E_{\alpha_0} \psi_0(x) + 2A\delta$$

and, by virtue of Step 5,

$$\int |F(x, \alpha_0) - \psi_0(x)| p(x, \alpha_0) dP_x \leq 4A\delta.$$

In consequence and on the strength of (16.41) we obtain

$$\int |F(x, \alpha_0) - \psi_0(x)| dP_x \leq \frac{4A\delta}{a} = \frac{\varepsilon}{4}.$$

This inequality holds true for any $\alpha_0 \in \Lambda_*^{**}$

In view of the triangle inequality and ε_0 -discreteness, it turns out that the class Λ_*^{**} consists of just one element, which contradicts property (a) of Step 4.

This completes the proof of Theorem 16.5a.

16.7 THE NECESSARY AND SUFFICIENT CONDITIONS

Theorem 16.6. *Let the conditions of measurability be satisfied for functions $F(x, a)$, $a \in A$. For one-sided uniform convergence to take place*

$$P \left\{ \sup_{\alpha \in \Lambda} \left(EF(x, \alpha) - \frac{1}{\ell} \sum_{i=1}^{\ell} F(x_i, a) \right) > \varepsilon_0 \right\} \xrightarrow[\ell \rightarrow \infty]{} 0,$$

it is necessary and sufficient that, for any positive numbers ϵ , η , and δ , there should exist a set of uniformly bounded measurable functions $\Phi(x, \alpha)$, $a \in A$, such that

$$\begin{aligned} F(x, \alpha) &\geq \Phi(x, \alpha), \\ E(F(x, a) - \Phi(x, a)) &< \delta, \end{aligned} \tag{16.42}$$

and for set $\Phi(x, a)$, $a \in A$, the inequality

$$C_\varepsilon^2 = \lim_{\ell \rightarrow \infty} \frac{H_\varepsilon^\Lambda(\ell)}{\ell} < \ln E + \eta, \tag{16.43}$$

is valid, where entropy $H_\varepsilon^\Lambda(\ell)$ is computed for the class $\Phi(x, a)$, $a \in A$.

Theorem 16.6a. *Let the class of densities $p(x, a)$, $\alpha \in A$, specifying the measure P_α with respect to the basis probability measure P , be such that the conditions of measurability are satisfied, and the densities themselves are uniformly bounded and uniformly separated from zero:*

$$0 < a \leq p(x, \alpha) \leq A < \infty.$$

Then, in order that the maximum likelihood method be nontrivial consistent for the class $p(x, a)$, $a \in A$, that is, for any $\alpha_0 \in A$

$$\inf_{\alpha \in \Lambda} \frac{1}{\ell} \sum_{i=1}^{\ell} (-\ln p(x_i, \alpha)) \xrightarrow[\ell \rightarrow \infty]{} E_{\alpha_0}(-\ln p(x, \alpha_0)),$$

it is necessary and sufficient that, for any positive numbers ϵ , η , and δ , there should exist a set of uniformly bounded functions $\Phi(x, \alpha)$, $\alpha \in A$, such that for any $\alpha \in A$

$$\begin{aligned} -\ln p(x, \alpha) &\geq \Phi(x, \alpha), \\ E_{\alpha_0}(-\ln p(x, \alpha) - \Phi(x, \alpha)) &< \delta \end{aligned} \quad (16.42a)$$

and for the set of functions $\Phi(x, \alpha)$, $\alpha \in A$ the inequality

$$C_\varepsilon^A = \lim_{\ell \rightarrow \infty} \frac{H_\varepsilon^\Lambda(\ell)}{\ell} < \ln \varepsilon + \eta \quad (16.43a)$$

is valid, where $H_\varepsilon^\Lambda(\ell)$ is computed for the class $\Phi(x, \alpha)$, $\alpha \in A$.

Proof of Sufficiency for Theorem 16.6. We assume the number ε_0 and evaluate the quantity

$$\begin{aligned} T_\ell &= P \left\{ \sup_{\alpha \in \Lambda} (EF(x, \alpha) - \frac{1}{\ell} \sum_{i=1}^{\ell} F(x_i, \alpha)) > \varepsilon_0 \right\} \\ &= \int \theta \left[\sup_{\alpha \in \Lambda} (EF(x, \alpha) - \frac{1}{\ell} \sum_{i=1}^{\ell} F(x_i, \alpha)) - \varepsilon_0 \right] dP_x. \end{aligned}$$

We set out to show that

$$\lim_{\ell \rightarrow \infty} T_\ell = 0.$$

We choose the positive numbers

$$\delta < \frac{\varepsilon_0}{2}, \quad \varepsilon = \frac{\varepsilon_0}{18}, \quad \eta = \frac{\varepsilon_0}{360}. \quad (16.44)$$

On their basis we find the class of functions $\Phi(x, \alpha)$, $\alpha \in A$, satisfying the conditions (16.42) and (16.43). For any $\alpha \in A$ we have

$$\begin{aligned} EF(x, \alpha) - \frac{1}{\ell} \sum_{i=1}^{\ell} F(x_i, \alpha) \\ \leq E\Phi(x, \alpha) + E(F(x, \alpha) - \Phi(x, \alpha)) - \frac{1}{\ell} \sum_{i=1}^{\ell} \Phi(x_i, \alpha) \\ \leq \left| E\Phi(x, \alpha) - \frac{1}{\ell} \sum_{i=1}^{\ell} \Phi(x_i, \alpha) \right| + \delta. \end{aligned}$$

Hence, subject to (16.44) we obtain

$$\begin{aligned} T_\ell &\leq \int \theta \left\{ \sup_{\alpha \in \Lambda} \left| E\Phi(x, \alpha) - \frac{1}{\ell} \sum_{i=1}^{\ell} \Phi(x_i, \alpha) \right| - (\varepsilon_0 + \delta) \right\} dP_x \\ &\leq P \left\{ \sup_{\alpha \in \Lambda} \left| E\Phi(x_i, \alpha) - \frac{1}{\ell} \sum_{i=1}^{\ell} \Phi(x_i, \alpha) \right| > \frac{\varepsilon_0}{2} \right\}. \end{aligned}$$

It was shown in Lemma 15.5 (see Chapter 15, Section 15.6) that for sufficiently large C_s and for $\varepsilon > 0$ we have

$$\begin{aligned} B_\ell &\equiv P \left\{ \sup_{\alpha \in \Lambda} \left| \frac{1}{\ell} \sum_{i=1}^{\ell} \Phi(x_i, \alpha) - \frac{1}{\ell} \sum_{i=\ell+1}^{2\ell} \Phi(x_i, \alpha) \right| > \varepsilon \right\} \\ &> \frac{1}{2} P \left\{ \sup_{\alpha \in \Lambda} \left| E\Phi(x_i, \alpha) - \frac{1}{\ell} \sum_{i=1}^{\ell} \Phi(x_i, \alpha) \right| > 3\varepsilon \right\}. \end{aligned}$$

In turn, as shown in Section 15.6, Eq. 15.39, for any $c > 0$ the following inequality holds true:

$$B_\ell \leq P \left\{ \frac{\ln N^\Lambda(\varepsilon/3; x_1, \dots, x_\ell)}{\ell} > c \right\} + 6\ell \exp \left[-\frac{\varepsilon^2(\ell+1)}{9} + cl \right].$$

By also setting $\varepsilon = \varepsilon_0/6$ and $c = \eta$, we obtain for sufficiently large ℓ 's

$$T_\ell \leq P \left\{ \frac{1}{\ell} \ln N^\Lambda \left(\frac{\varepsilon_0}{18}; x_1, \dots, x_\ell \right) > \eta \right\} + 6\ell \exp \left[- \left(\frac{\varepsilon_0^2}{360} - \eta \right) / \ell \right]. \quad (16.45)$$

Now it follows from (16.3) that

$$P \left\{ \frac{\ln N^\Lambda(\varepsilon; x_1, \dots, x_\ell)}{\ell} > \eta \right\} < P \left\{ \frac{\ln V_\varepsilon(x_1, \dots, x_\ell)}{\ell} > \ln \varepsilon + \eta \right\}.$$

Finally, we obtain from (16.2), (16.43), (16.44), and (16.45)

$$T_\ell = P \left\{ \sup_{\alpha \in \Lambda} \left(EF(x, \alpha) - \frac{1}{\ell} \sum_{i=1}^{\ell} F(x_i, \alpha) \right) > \varepsilon \right\} \xrightarrow{\ell \rightarrow \infty} 0.$$

The sufficiency has thus been proved.

Proof of Sufficiency for Theorem 16.6a. The consistency of the maximum likelihood method for a class of densities $p(x, a)$, $a \in A$, follows from the one-sided uniform convergence for the class of functions $-\ln p(x, a)$ in any

specified measure P_{α_0} , $\alpha_0 \in A$. To demonstrate, we set $\alpha_0 \in A$. Then, by the law of large numbers we have

$$\frac{1}{\ell} \sum_{i=1}^{\ell} -\ln p(x, \alpha_0) \xrightarrow[\ell \rightarrow \infty]{P_{\alpha_0}} E_{\alpha_0} - \ln p(x, \alpha_0).$$

Therefore it will suffice to establish that for any $\varepsilon > 0$ we obtain

$$P \{ E_{\alpha_0} (-\ln p(x, q)) - \inf_{\alpha \in A} \frac{1}{\ell} \sum_{i=1}^{\ell} (-\ln p(x_i, \alpha)) > \varepsilon \} \xrightarrow[\ell \rightarrow \infty]{} 0.$$

But for any $a \in A$ we have

$$E_{\alpha_0} \ln p(x, a) \leq E_{\alpha_0} \ln p(x, \alpha_0).$$

Therefore the statement of Theorem 16.6a follows from the condition

$$P \left\{ \sup_{\alpha} \left[E_{\alpha_0} (-\ln p(x, a)) - \frac{1}{\ell} \sum_{i=1}^{\ell} (-\ln p(x_i, a)) \right] > \varepsilon \right\} \xrightarrow[\ell \rightarrow \infty]{} 0$$

for any $\varepsilon > 0$, that is, from condition for one-side uniform convergence.

From Corollary 1 of Theorem 16.3 it follows, however, that if the conditions of Theorem 16.6a hold for the measure P , then they will hold for any measure P_{α} as well. By applying Theorem 16.6 (sufficiency) we obtain the sought-after result.

Proof of Necessity. For purposes of proof, we will need one more lemma.

Lemma 16.11. Let $F(x, a)$, $a \in A$, be measurable in x and uniformly bounded functions defined on X . Furthermore, let $B \subset X$ and

- (a) for any $\delta > 0$, $k > 0$ and for almost any sample x_1, \dots, x_k ($x_i \in B$), let there exist $a^* \in A$ such that

$$|F(x_i, a^*) - \psi_0(x_i)| < \delta, \quad (16.46)$$

and

- (b) for some $\delta_0 > 0$ let

$$P \left\{ \sup_{\alpha \in A} \left(EF(x, \alpha) - \frac{1}{\ell} \sum_{i=1}^{\ell} F(x_i, \alpha) \right) > \delta_0 \right\} \xrightarrow[\ell \rightarrow \infty]{} 0.$$

Then

$$\int_B \psi_0(x) dP_x \geq \inf_{\alpha \in A} \int_B F(x, a) dP_x - \delta.$$

Proof. Note that with $P(B) = 0$ the lemma is trivial.

Suppose that $P(B) \geq 0$ and suppose that conclusion of the lemma is incorrect. Then there should exist a number $\varepsilon_0 > 0$ such that

$$\int_B \psi_0(x) dP_x < \inf_{\alpha \in \Lambda} \int_B F(x, \alpha) dP_x + (\varepsilon_0 - \delta). \quad (16.47)$$

To begin with, we consider the case $P(B) = 1$. We denote

$$S_\ell = \int_B \theta \left\{ \sup_{\alpha \in \Lambda} \left(EF(x, \alpha) - \frac{1}{\ell} \sum_{i=1}^{\ell} F(x_i, \alpha) \right) - \delta_0 \right\} dP_x.$$

Condition (b) implies that

$$\lim_{\ell \rightarrow \infty} S_\ell = 0.$$

We fix the sample x_1, \dots, x_ℓ and, in view of condition (a), we choose a^* such that

$$|F(x_i, a^*) - \psi_0(x)| < \frac{\varepsilon_0}{2}.$$

This leaves us with a string of inequalities that hold true for almost any sample:

$$\begin{aligned} \phi &\equiv \sup_{\alpha \in \Lambda} \left(EF(x, \alpha) - \frac{1}{\ell} \sum_{i=1}^{\ell} F(x_i, \alpha) \right) - \delta_0 \\ &\geq EF(x, a^*) - \frac{1}{\ell} \sum_{i=1}^{\ell} F(x_i, a^*) - \delta_0 \\ &\geq EF(x, a^*) - \frac{1}{\ell} \sum_{i=1}^{\ell} \psi_0(x_i) - \left(\delta_0 - \frac{\varepsilon_0}{2} \right) \\ &\geq \inf_{\alpha \in \Lambda} EF(x, \alpha) - \frac{1}{\ell} \sum_{i=1}^{\ell} \psi_0(x_i) - \left(\delta_0 + \frac{\varepsilon_0}{2} \right). \end{aligned}$$

By applying (16.47) and noting that $P(B) = 1$, we obtain

$$\begin{aligned} \phi &\geq \int_B \psi_0(x) dP_x + (\delta + \varepsilon_0) - \frac{1}{\ell} \sum_{i=1}^{\ell} \psi_0(x_i) - \left(\delta + \frac{\varepsilon_0}{2} \right) \\ &= \int_B \psi_0(x) dP_x - \frac{1}{\ell} \sum_{i=1}^{\ell} \psi_0(x_i) + \frac{\varepsilon_0}{2}. \end{aligned}$$

Going back to estimate S_ℓ , we get

$$S_\ell \geq P \left\{ E\psi_0(x) - \frac{1}{\ell} \sum_{i=1}^{\ell} \psi_0(x_i) > -\frac{\varepsilon_0}{2} \right\}.$$

By the law of large numbers, however, the right-hand side of the inequality tends to one, and this contradicts condition (b).

We now pass to the case $0 < P(B) < 1$. It is an easy matter to see that for an arbitrary symmetric function $g(x_1, \dots, x_\ell)$, where x_1, \dots, x_ℓ is the sample collected in a series of independent trials with a distribution P_x and for an arbitrary set $B \subset X$, the following identity holds:

$$\begin{aligned} & \int g(x_1, \dots, x_\ell) dP_{x_1, \dots, x_\ell} \\ &= \sum_{k=0}^{\ell} C_\ell^k \int_{x_1, \dots, x_k \in B} \left[\int_{x_{k+1}, \dots, x_\ell \notin B} g(x_1, \dots, x_\ell) dP_{x_{k+1}, \dots, x_\ell} \right] dP_{x_1, \dots, x_k}. \end{aligned}$$

We now introduce on X the densities

$$\begin{aligned} \pi_1(x) &= \begin{cases} \frac{1}{P(B)} & \text{if } x \in B, \\ 0 & \text{if } x \notin B, \end{cases} \\ \pi_2(x) &= \begin{cases} 0 & \text{if } x \in B, \\ \frac{1}{P(B)} & \text{if } x \notin B, \end{cases} \end{aligned}$$

and denote the measures $P^{(1)}$ and $P^{(2)}$ defined by the conditions

$$dP^{(1)} = \pi_1 dP_x, \quad dP^{(2)} = \pi_2 dP_x.$$

Then

$$\begin{aligned} \int g(x_1, \dots, x_\ell) dP_{x_1, \dots, x_\ell} &= \sum_{k=0}^{\ell} C_\ell^k P^k(B)(1 - P(B))^{\ell-k} \\ &\quad \times \int_{x_1, \dots, x_k} \left[\int_{x_{k+1}, \dots, x_\ell} g(x_1, \dots, x_\ell) dP_{x_{k+1}, \dots, x_\ell}^{(2)} \right] dP_{x_1, \dots, x_k}^{(1)}. \end{aligned} \tag{16.48}$$

We denote now

$$\begin{aligned} S_\ell &= P \left\{ \sup_{\alpha \in \Lambda} \left(EF(x, \alpha) - \frac{1}{\ell} \sum_{i=1}^{\ell} F(x_i, \alpha) \right) > \delta_0 \right\} \\ &= \int \theta \left[\sup_{\alpha \in \Lambda} \left(EF(x, \alpha) - \frac{1}{\ell} \sum_{i=1}^{\ell} F(x_i, \alpha) \right) - \delta_0 \right] dP_x. \end{aligned} \tag{16.49}$$

By condition (b),

$$\lim_{\ell \rightarrow \infty} S_\ell = 0.$$

We will let this requirement contradict the assumption (16.47).

We denote the function under the integral sign as $g(x_1, \dots, x_\ell)$ and use relation (16.48). We set

$$\sup_{\alpha, x} F(x, \alpha) = A_0.$$

We fix ℓ , k , and a part of the sample x_1, \dots, x_k , assuming that the following conditions are satisfied:

$$\begin{aligned} \left| 1 - \frac{k}{\ell P(B)} \right| &< \frac{\varepsilon_0}{8A_0}, \\ \left| 1 - \frac{\ell - k}{\ell(1 - P(B))} \right| &< \frac{\varepsilon_0}{8A_0}, \\ \left| 1 - \frac{k}{\ell P(B)} \right| &< \frac{\varepsilon_0}{8(\varepsilon_0 + \delta_0)}, \end{aligned} \quad (16.50)$$

and also $x_i \in B$ ($1 \leq i \leq k$):

$$\left| E_1 \psi_0(x) - \frac{1}{k} \sum_{i=1}^k \psi_0(x_i) \right| < \frac{\varepsilon_0}{8}, \quad (16.51)$$

where

$$E_1 \psi_0(x) = \int \psi_0(x) dP_x^{(1)} = \frac{1}{P(B)} \int_B \psi_0(x) dP_x$$

Using condition (a), we choose α^* so as to satisfy the condition

$$|\psi_0(x_i) - F(x_i, \alpha^*)| < \frac{\varepsilon_0}{8}. \quad (16.52)$$

Now we have

$$\begin{aligned} \phi &\equiv \sup_{\alpha \in \Lambda} \left(EF(x, \alpha) - \frac{1}{\ell} \sum_{i=1}^{\ell} F(x_i, \alpha) \right) \\ &\geq EF(x, \alpha^*) - \frac{1}{\ell} \sum_{i=1}^{\ell} F(x_i, \alpha^*). \end{aligned}$$

We denote

$$\begin{aligned} E_B F(x, \alpha) &= \int_B F(x, \alpha) dP_x, \\ E_{\bar{B}} F(x, \alpha) &= \int_{\bar{B}} F(x, \alpha) dP_x, \end{aligned}$$

such that

$$EF(x, \alpha) = E_B F(x, \alpha) + E_{\bar{B}} F(x, \alpha),$$

$$E_B F(x, \alpha) \leq A_0, \quad E_{\bar{B}} F(x, \alpha) \leq A_0.$$

Furthermore, the following identities hold:

$$\begin{aligned} EF(x, \alpha^*) &= \frac{1}{\ell} \sum_{i=l}^{\ell} F(x_i, \alpha^*) \\ E_B F(x, \alpha^*) + E_{\bar{B}} F(x, \alpha^*) &= \frac{1}{\ell} \left[\sum_{i=l}^k F(x_i, \alpha^*) + \sum_{i=k+1}^{\ell} F(x_i, \alpha^*) \right] \\ &= \frac{k}{\ell} \left[E_1 F(x, \alpha^*) - \frac{1}{k} \sum_{i=1}^k F(x_i, \alpha^*) \right] \\ &\quad + \frac{\ell-k}{\ell} \left[E_2 F(x, \alpha^*) - \frac{1}{\ell-k} \sum_{i=k+1}^{\ell} F(x_i, \alpha^*) \right] \\ &\quad + E_B F(x, \alpha^*) \left(1 - \frac{k}{\ell P(B)} \right) + E_{\bar{B}} F(x, \alpha^*) \left(1 - \frac{\ell-k}{\ell(1-P(B))} \right), \end{aligned} \quad (16.53)$$

where

$$E_1 F(x, \alpha) = \int F(x, \alpha) dP_x^{(1)} = \frac{E_B F(x, \alpha)}{P(B)},$$

$$E_2 F(x, \alpha) = \int F(x, \alpha) dP_x^{(2)} = \frac{E_{\bar{B}} F(x, \alpha)}{1-P(B)}.$$

By denoting as T the quantity

$$T = E_1 F(x, \alpha^*) - \frac{1}{k} \sum_{i=1}^k F(x_i, \alpha)$$

we have, by virtue of (16.51) and (16.52),

$$\begin{aligned} T &\geq E_1 F(x, \alpha^*) - \frac{1}{k} \sum_{i=1}^k \psi_0(x_i) - \frac{\varepsilon_0}{8} \\ &\geq (E_1 F(x, \alpha^*) - E_1 \psi_0(x)) + \left(E_1 \psi_0(x) - \frac{1}{k} \sum_{i=1}^k \psi_0(x_i) \right) - \frac{\varepsilon_0}{8} \\ &\geq E_1 F(x, \alpha^*) - E_1 \psi_0(x) - \frac{\varepsilon_0}{4} \\ &\quad - \frac{1}{P(B)} \int (F(x, \alpha^*) - \psi_0(x)) dP_x - \frac{\varepsilon_0}{4}. \end{aligned}$$

Now it follows from (16.47) that

$$\begin{aligned} T &\geq \frac{1}{P(B)} \left[\int_B F(x, \alpha^*) dP_x - \int_B \psi_0(x) dP_x \right] - \frac{\varepsilon_0}{4} \\ &\geq \frac{1}{P(B)} \left[\inf_{\alpha \in \Lambda} \int_B F(x, \alpha) dP_x - \int_B \psi_0(x) dP_x \right] - \frac{\varepsilon_0}{4} \\ &> \frac{1}{P(B)} (\delta_0 + \varepsilon_0) - \frac{\varepsilon_0}{4}. \end{aligned}$$

And further from (16.50) we have

$$\begin{aligned} \frac{k}{\ell} T &\geq \frac{k}{\ell P(B)} (\delta_0 + \varepsilon_0) - \frac{\varepsilon_0}{4} \\ &= (\delta_0 + \varepsilon_0) - \left(1 - \frac{k}{\ell P(B)}\right) (\delta_0 + \varepsilon_0) - \frac{\varepsilon_0}{4} \\ &\geq \delta_0 + \frac{5}{8} \varepsilon_0. \end{aligned}$$

Going back to the estimate ϕ , we obtain from (16.53)

$$\begin{aligned} \mathbf{S} &\geq \delta_0 + \frac{5}{8} \varepsilon_0 + \frac{\ell - k}{\ell} \left[E_2 F(x, \alpha^*) - \frac{1}{e - k} \sum_{i=k+1}^{\ell} F(x_i, \alpha^*) \right] \\ &\quad + E_B F(x, \alpha^*) \left(1 - \frac{k}{\ell P(B)}\right) + E_{\bar{B}} F(x, \alpha^*) \left(1 - \frac{\ell - k}{\ell(1 - P(B))}\right). \end{aligned}$$

By applying (16.50), we obtain

$$\phi \geq \frac{3}{8} \varepsilon_0 + \delta_0 - \left| E_2 F(x, \alpha^*) - \frac{1}{e - k} \sum_{i=k+1}^{\ell} F(x_i, \alpha^*) \right|.$$

Thus, for ℓ , k , and x_1, \dots, x_k satisfying conditions (16.50) and (16.51) we have

$$g(x_1, \dots, x_\ell) = \theta[\phi - \delta_0] \geq \theta \left(\frac{3}{8} \varepsilon_0 - \left| E_2 F(x, \alpha^*) - \frac{1}{e - k} \sum_{i=k+1}^{\ell} F(x_i, \alpha^*) \right| \right)$$

and

$$\begin{aligned} \int g(x_1, \dots, x_\ell) dP_{x_{k+1}, \dots, x_\ell}^{(2)} \\ \geq P \left\{ \left| E_2 F(x, \alpha^*) - \frac{1}{e - k} \sum_{i=k+1}^{\ell} F(x_i, \alpha^*) \right| \leq \frac{3}{8} \varepsilon_0 \right\}. \end{aligned}$$

By the law of large numbers for uniformly bounded functions the last expression tends to unity as $(p - k)$ tends to infinity with an estimate which solely depends on $\ell - k$. That is, there exists an estimate

$$\rho_1(\ell - k) \leq P \left\{ \left| E_2 F(x, \alpha^*) - \frac{1}{\ell - k} \sum_{i=k+1}^{\ell} F(x_i, \alpha^*) \right| \leq \frac{3}{8} \varepsilon_0 \right\}$$

such that

$$\lim_{(\ell-k) \rightarrow \infty} \rho_1(\ell - k) = 1.$$

Thus, in conditions (16.50) and (16.51) we have

$$R(x_1, \dots, x_k) = \int_{x_{k+1}, \dots, x_\ell} g(x_1, \dots, x_\ell) dP_{x_{k+1}, \dots, x_\ell}^{(2)} \geq \rho_1(\ell - k).$$

Going back to the estimate S_ℓ (see (16.49)) and using (16.48), we get

$$\begin{aligned} S_\ell &= \sum_{k=0} C_\ell^k P^k(B) (1 - P(B))^{\ell-k} \int_{x_1, \dots, x_\ell} R(x_1, \dots, x_k) dP_{x_1, \dots, x_k}^{(1)} \\ &\geq \sum^* C_\ell^k P^k(B) (1 - P(B))^{\ell-k} \rho_1(\ell - k) \int_{\hat{X}} dP_{x_1, \dots, x_k}, \end{aligned}$$

where \sum^* is taken only over k 's satisfying (16.50). and X is the set of sequences x_1, \dots, x_k satisfying (16.51). From (16.51) we have

$$\int_{\hat{X}} dP_{x_1, \dots, x_k} = P \left\{ \left| E_1 \psi_0(x) - \frac{1}{k} \sum_{i=1}^k \psi_0(x_i) \right| \leq \frac{\varepsilon_0}{8} \right\}.$$

By the law of large numbers, the last expression for a bounded quantity tends to unity as k increases, with an estimate depending solely on k , that is,

$$\int_{\hat{X}} dP_{x_1, \dots, x_k} \geq \rho_2(k), \quad \lim_{k \rightarrow \infty} \rho_2(k) = 1$$

By extending the estimate S_ℓ , we obtain

$$S_\ell \geq \sum^* C_\ell^k P^k(B) (1 - P(B))^{\ell-k} \rho_1(p - k) \rho_2(k).$$

Note that with ℓ tending to infinity, all k 's and $(p - k)$'s satisfying (16.50) uniformly tend to infinity. Hence

$$\lim_{\ell \rightarrow \infty} S_\ell \geq \lim_{\ell \rightarrow \infty} \sum_k C_\ell^k P^k(B) (1 - P(B))^{\ell-k}.$$

By the law of large numbers for the binomial distribution we obtain

$$\lim_{\ell \rightarrow \infty} S_\ell = 1,$$

in contradiction to condition (b). The lemma has thus been proved.

Continued Proof of Theorem 15.6 (Necessity). We propose the following method for constructing the class of functions $\Phi(x, a)$, $a \in A$. Let the positive numbers ε , S , and η be specified. We set

$$\delta_0 = \min \left\{ \frac{\varepsilon \eta}{4}, \delta \right\} \quad (16.54)$$

By Theorem 16.5, the class $F(x, a)$, $a \in A$, can be decomposed into a finite number of subclasses Λ_i so that the diameter of each in $L_1(P)$ is smaller than δ_0 , that is,

$$\sup_{\alpha_1, \alpha_2 \in \Lambda_i} \int |F(x, a) - F(x, \alpha_2)| dP_x < \delta_0. \quad (16.55)$$

In each subclass, we select one function $F(x, a_i)$, $a_i \in \Lambda_i$. To each function $F(x, a)$, $a \in A$, there corresponds a new function

$$\Phi(x, a) = \min(F(x, a), F(x, a_i)).$$

The class of all functions $\Phi(x, a)$, $a \in A$, is the one sought.

Indeed conditions (16.42) stem immediately from (16.54) and (16.55) and from the definition of $\Phi(x, a)$. Only (16.43) remains to be checked.

Suppose that (16.43) is not satisfied; that is, for the class $\Phi(x, a)$, $a \in A$, we have

$$C_\varepsilon^\Lambda \geq \ln \varepsilon + \eta.$$

Then for at least one subclass $\Phi(x, \alpha)$, $a \in \Lambda_i$, we obtain

$$C_\varepsilon^{\Lambda_i} \geq \ln \varepsilon + \eta.$$

We fix Λ_i and let the assumption contradict (16.53). By Theorem 16.4, there exist such functions $\psi_1(x)$ and $\psi_0(x)$ that

- (a) $\psi_1(x) \geq \psi_0(x)$;
- (b) $\int |\psi_1(x) - \psi_0(x)| dP_x \geq \exp C_\varepsilon^{\Lambda_i} - \varepsilon \geq (e^\eta - 1)\varepsilon \geq \eta\varepsilon$;
- (c) $F(x, \alpha_i) = \max_{\alpha \in \Lambda_i} \Phi(x, \alpha) \geq \psi_1(x)$;
- (d) for almost any sequence x_1, \dots, x_ℓ and for any number $a > 0$, one can find $a^* \in A^*$ such that

$$|\Phi(x_j, a^*) - \psi_0(x_j)| < \sigma, \quad j = 1, \dots, \ell.$$

In view of (b) and (c), we have

$$\int (F(x, \alpha_i) - \psi_0(x)) dP_x > \int |\psi_1(x) - \psi_0(x)| dP_x > \varepsilon\eta. \quad (16.56)$$

We denote by $B \subset X$ the set

$$\{x : \psi_0(x) < F(x, \alpha_i) - \delta\}.$$

Then for almost any sequence $x_1, \dots, x_\ell \in B$ and for any number $\sigma > 0$ there will be $a^* \in A$ such that

$$|F(x_j, a^*) - \psi_0(x_j)| < \sigma, \quad j = 1, \dots, \ell. \quad (16.57)$$

For this to happen, it suffices to choose a positive

$$\sigma^* < \min(\sigma, \delta_0)$$

and, by taking advantage of (d), to find a^* satisfying

$$|\Phi(x_j, a^*) - \psi_0(x_j)| < \sigma^*, \quad j = 1, \dots, \ell.$$

Furthermore, since all $x_j \in B$ we have

$$\Phi(x_j, a^*) < \psi_0(x_j) + \sigma^* < \psi_0(x_j) + \delta_0 < F(x_j, \alpha_i).$$

Therefore taking into account the definition of $\Phi(x, a^*)$ we conclude that

$$\Phi(x_j, a^*) = F(x_j, a^*).$$

Now we will apply Lemma 16.11 to the subset Λ_i (condition (a) of Lemma 16.11 has just been shown, and condition (b) follows from one-sided uniform convergence). By virtue of its conclusion we obtain

$$\int_B \psi_0(x) dP_x \geq \inf_{a \in \Lambda_i} \int_B F(x, a) dP_x - \delta. \quad (16.58)$$

From (16.56) we have

$$\begin{aligned} \varepsilon\eta &< \int (F(x, \alpha_i) - \psi_0(x)) dP_x \\ &= \int_B (F(x, \alpha_i) - \psi_0(x)) dP_x + \int_{\bar{B}} (F(x, \alpha_i) - \psi_0(x)) dP_x. \end{aligned}$$

On the set B it is true that

$$F(x, \alpha_i) - \psi_0(x) < \delta_0.$$

Therefore it follows from (16.58) that

$$\begin{aligned}\varepsilon\eta &< \delta_0 + \int_B F(x, \alpha_i) dP_x - \int_B \psi_0(x) dP_x \\ &\leq 2\delta_0 + \int F(x, \alpha_i) dP_x - \inf_{\alpha \in \Lambda_i} \int F(x, \alpha) dP_x\end{aligned}$$

and from (16.54) we have

$$\varepsilon\eta < 2\delta_0 + \sup_{\alpha_1, \alpha_2 \in \Lambda_i} |F(x, \alpha_1) - F(x, \alpha_2)| < 3\delta_0,$$

in contradiction to (16.53).

Thus Theorem 16.6 has been proved.

Proof of Necessity for Theorem 16.6a. By conditions of Theorem 16.6a, for any $\alpha_0 \in A$ we have

$$a\rho(\alpha_1, \alpha_2) \leq \int |\ln p(x, \alpha_1) - \ln p(x, \alpha_2)| dP_{\alpha_0} \leq A\rho(\alpha_1, \alpha_2), \quad (16.59)$$

where

$$\rho(\alpha_1, \alpha_2) = \int |\ln p(x, \alpha_1) - \ln p(x, \alpha_2)| dP_x.$$

For the specified positive ε, δ , and η , we set

$$< \min\left(\frac{\varepsilon\eta a}{2(A+1)}, a\right). \quad (16.60)$$

By (16.59), the set A can be decomposed into a finite number of subclasses A_i , such that

$$\rho(\alpha_1, \alpha_2) < \delta_0, \quad \alpha_1, \alpha_2 \in A_i.$$

In each subclass we select $a_i \in A_i$, and put

$$\Phi(x, \alpha) = \min(-\ln p(x, \alpha), -\ln p(x, \alpha_i)), \quad \alpha \in A_i.$$

Conditions (16.42a) follow immediately from the definition. Suppose that (16.43a) is not satisfied, that is, for at least one subclass $\Phi(x, \alpha), \alpha \in A_i$, one has

$$C_\varepsilon^{\Lambda_i} \geq \ln \varepsilon + \eta.$$

On denoting

$$F(x, \alpha) = -\ln p(x, \alpha)$$

we can, as in the proof of Theorem 16.6, choose functions $\psi_1(\mathbf{x})$ and $\psi_0(x)$ satisfying conditions (a), (b), (c), and (d) and (16.56). The property (16.57) will likewise be satisfied; furthermore, on fixing $\mathbf{a} \in A$, it may be taken that it is satisfied for almost any sequence in the sense of P_α .

In order to be able to apply Lemma 16.11, however, we now have to resort to the nontrivial consistency of the maximum likelihood method, instead of one-sided uniform convergence. Let us apply this condition to the case $\alpha_0 = \mathbf{a}$:

$$\inf_{\alpha \in \Lambda_\ell} \frac{1}{\ell} \sum_{j=1}^{\ell} F(x_j, \alpha) \xrightarrow[\ell \rightarrow \infty]{P_{\alpha_0}} E_{\alpha_0} F(x, \mathbf{a}).$$

Therefore

$$P \left\{ E_{\alpha_0} F(x, \alpha_i) - \frac{1}{\ell} \sum_{j=1}^{\ell} F(x_j, \alpha_i) > \delta_0 \right\} \xrightarrow[\ell \rightarrow \infty]{} 0. \quad (16.61)$$

In view of (16.59) and of the choice of A , we have

$$\sup_{\alpha \in \Lambda_\ell} |E_{\alpha_0} F(x, \alpha) - E_{\alpha_0} F(x, \alpha_i)| < A \delta_0.$$

Hence

$$\sup_{\alpha \in \Lambda_\ell} |E_{\alpha_0} F(x, \alpha) - E_{\alpha_0} F(x, \alpha_i)| < A \delta_0.$$

By combining the above inequality with (16.51), we obtain

$$P \left\{ \sup_{\alpha \in \Lambda_\ell} \left(E_{\alpha_0} F(x, \alpha) - \frac{1}{\ell} \sum_{j=1}^{\ell} F(x_j, \alpha) \right) > (A+1) \delta_0 \right\} \xrightarrow[\ell \rightarrow \infty]{} 0.$$

We now apply Lemma 16.11 to the class $\Phi(x, \mathbf{a})$ for the measure P_{α_0} . For any measurable $B \subset X$ we get

$$\int \psi_0(x) dP_{\alpha_0} \geq \inf_{\alpha \in \Lambda_\ell} \int_B F(x, \alpha) dP_{\alpha_0} - (A+1) \delta_0. \quad (16.62)$$

From (16.56) and (16.59) we have

$$\varepsilon \eta < \int (F(x, \alpha_i) - \psi_0(x)) dP_x \leq \frac{1}{a} \int (F(x, \alpha_i) - \psi_0(x)) dP_{\alpha_0}.$$

As before, we denote

$$B = \{x : \psi_0(x) < F(x, \alpha_i) - \delta_0\}.$$

From (16.62) we obtain

$$\begin{aligned}\varepsilon \eta &< \frac{1}{a} \left[\int_B (F(x, \alpha_i) - \psi_0(x)) dP_{\alpha_i} + \int_{\hat{B}} (F(x, \alpha_i) - \psi_0(x)) dP_{\alpha_i} \right] \\ &\leq \frac{1}{a} \left[\delta_0 + \delta_0(A+1) + \int_B F(x, \alpha_i) dP_{\alpha_i} - \inf_{\alpha \in \Lambda_i} \int_B F(x, \alpha_i) dP_{\alpha_i} \right].\end{aligned}$$

And, finally,

$$\varepsilon \eta < \frac{1}{a} \left[\delta_0(A+2) + \sup_{\alpha_1, \alpha_2 \in \Lambda_i} \int |F(x, \alpha_1) - F(x, \alpha_2)| dP_{\alpha_i} \right] \leq \frac{2\delta_0(A+1)}{a},$$

in contradiction to (16.60).

This completes the proof of Theorem 16.6a.

COMMENTS AND BIBLIOGRAPHICAL REMARKS

INTRODUCTION

Two events that transformed the world forever occurred during the twentieth century: scientific progress and the information technology revolution.

Modern scientific progress began at the turn of this century. It changed our philosophy and our understanding of general models of the world, shifting them from purely deterministic to stochastic. Fifty years later the information technology revolution began. It had enormous impact on life in general, opening new opportunities and enabling people to be more creative in solving everyday tasks.

In discussing scientific progress, one usually considers physics as the primary example of changing the general model of the world in a relatively short time: from Newton's macro-models of relatively small velocity to micro-models of quantum mechanics and physics of high velocities (the theory of relativity). It is often stressed that at the time these models were introduced they were not considered as something of practical importance. The history of physics collected a number of remarks in which creators of new models were skeptical of practical use for their theories. In 50 years, however, new theories and new ways of thinking became the basis for a technological revolution. As we will see this underestimation of theoretical models was not only specific to physics.

Revolutionary changes took place in many branches of science. For us it is important that new ideas also occurred in understanding the principles of inductive inference and creating statistical methods of inference. The names of three great scientists who addressed the problem of induction from different points of view should be mentioned with regard to these new ideas:

Karl Popper, who considered the problem of induction from a philosophical perspective;

Andrei N. Kolmogorov, who considered the problem of induction from a statistics foundation perspective;

Ronald A. Fisher, who considered the problem of induction from an applied statistics perspective.

The problem of inductive inference has been known for more than two thousand years. In the Introduction we discussed the Occam razor principle dating back to the fourteenth century. The modern analysis of induction, as a problem of deriving universal assertions from particular observations, was started, however, in the eighteenth century when D. Hume and especially I. Kant introduced the problem of demarcation:

What is the distinction between empirical theories (theories that reflect truth for our world) and mathematics, logic, and metaphysical systems?

In the 1930s K. Popper proposed his solution to the demarcation problem based on the concept of falsifiability of theories. He proposed to consider as a necessary condition for correctness of empirical theories the possibility of their falsification. The easier it is to falsify a theory, the better the chances that the theory is true. In his solution of the demarcation problem, K. Popper for the first time connected the generalization ability with the capacity concept. His demarcation principle was very general. It was not restricted to some specific mathematical model in the framework of which one could provide exact analysis. Nevertheless, it describes one of the main factors contributing to generalization (the capacity factor) that will later appear as a result of exact analysis in statistical learning theory.

At approximately the same time as Popper, Glivenko and Cantelli proved that an empirical distribution function converges to the actual one with increasing number of observations and Kolmogorov found the asymptotically exact rate of convergence. These results demonstrated that one could find an approximation to a distribution function that is close to the desired one in the metric C. Of course these results still were far from solving the main problem of statistics: estimating the unknown probability measure. To estimate a probability measure we need convergence of estimates to the actual distribution function in a metric that is stronger than C. Nevertheless, existence of a fast (exponential) rate of convergence of the empirical distribution function to the actual one gave a hope that this problem had a solution.

These two results—namely, the discovery of the factors responsible for generalization and the discovery of the first (asymptotically exact) bound on the rate of uniform convergence of the frequencies to their probabilities for the special set of events—were a good start in developing general methods of statistical inductive inference. It was clear that as soon as a general theory of estimating the probability measure was developed, it would be possible to construct general statistical methods of induction useful for practical applications.

However, a particular approach was suggested by R. Fisher approximately at the same time when K. Popper and A. N. Kolmogorov made the very first steps toward the general theory of inductive inference.

R. Fisher, who was involved in many applied projects needed statistical inference not in 20 years, but immediately. Moreover, he needed methods based on simple calculations. Under these restrictions he suggested an excellent solution. He simplified the core problem of statistical inference—estimating probability measures—by reducing it to the problem of estimating parameters of density function. Then he developed, on the basis of this simplified core problem, almost all branches of modern statistics such as discriminant analysis, regression analysis, and density estimation.

Fisher's simplification of the core problem of statistical inference and his success in solving simple practical problems had deep consequences. It split statistical science into two parts: theoretical statistics, the branch of science that considers general methods of inference, and applied statistics, the branch of science that considers particular models of inference.

Due to the excellent development of the simplified approach, it became common opinion that for practical purposes the simplified version of statistical inference is sufficient and theoretical statistics was not considered an important source for new ideas in inductive inference.

As soon as the information technology revolution provided opportunities for estimating high-dimensional functions (the 1960s), the "curse of dimensionality" was discovered: that is the difficulties that arise when one considers multi-dimensional problems. In fact it was discovered that it is impossible to beat the curse of dimensionality in the framework of Fisher's paradigm. It should be noted that belief in this paradigm was so strong that for more than 25 years nobody tried to overcome it.[†] The curse of dimensionality was accepted as a law of nature that should be taken into account when solving problems.

To overcome this curse, one should come back to the theoretical foundation of statistics in order to identify factors responsible for generalization which in many ways reflected (a) the philosophy discussed in the 1930s by K. Popper and (b) the analysis of uniform convergence of frequencies to their probabilities that was started for the particular cases by Glivenko, Cantelli, and Kolmogorov.

CHAPTER 1

The Beginning

Now it is hard to say who made the very first step toward Statistical Learning Theory which suggested that we consider the problem of minimization of the risk functional based on empirical data as a general learning problem. For me

[†]The idea that it is possible to overcome the curse of dimensionality using a neural network was expressed for the first time in the early 1990s.

this happened in 1963 at machine learning seminars at the Moscow Institute of Control Sciences.

That year, at one of the seminars Novikoff's theorem on convergence of the perceptron algorithm was discussed. This theorem had tremendous impact on the audience.

Nowadays, it is hard to understand why this theorem, whose proof is based on simple arguments, could make such a big impact. The explanation is probably the following: The early 1960s were the beginning of the revolution in information technology. In the following pages, we will see, even focusing on one specific area, how many new ideas were originated at this time.

In particular, in the beginning of the 1960s, for the first time, people associated their future with the computer revolution. They tried to understand future technology, its influence on human values, and its impact on the development of science.

These discussions were started by the famous book entitled *Cybernetics* (the new word invented for this new subject), written by a remarkable mathematician N. Wiener. Wiener tried to describe his vision of future involvement of mathematical methods in everyday life where, by using computers, one would be able to solve intellectual tasks. This book was a great success. After Wiener's book, there appeared a number of publications written by specialists in different areas who described their visions of future computer civilizations.

Most of these books, however, saw the source of success in solving intellectual problems in the power of computers rather than in the power of mathematical analysis. It created the impression that exact mathematical analysis of intellectual problems was the old-fashioned way of solving them. Using computers and simple algorithms that imitate methods used by people (or animals or nature), one could achieve the highest level of performance simply due to the power of computers. The problem of imitating is not very complicated: It is enough to observe carefully the way in which the solution was obtained by humans and describe it as an algorithm.

The very first experiments with toy problems demonstrated the first success of this philosophy. In the early 1960s it looked as if the next step would bring significant results. The next step has never come to be.

Nevertheless, computer hardliners declared the creed, reiterated even in the late 1990s:

Complex theories do not work, simple algorithms do.

It should be noted that this declaration was not immediately rejected by scientific community[†]: Many positive revolutionary changes occurred in the 1960s, and it was not clear if this declaration was not one of them. The

[†]It is not easy to reject this philosophy. For example, should one of the last ideas of this type, the genetic programming still popular in the late 1990s, be rejected*?

reaction of a large part of the scientific community somehow reflected this philosophy: The interest shifted to the problem of organizations and construction of complex systems based on primitive automata. People were ready to accept the philosophy according to which it is not necessary to understand what is going on (what are the general principles) but is, instead, enough to understand how it is going on (how these principles are implemented). Therefore it became popular to describe complex behavior as a result of primitive actions of a large number of simple agents and to associate complexity of behavior with the number of agents. To specify what type of simple agents to use and in what kind of simple interaction they are involved, one had to analyze human solutions of the problems.

Therefore the main researchers in cybernetics became biologists, physiologists, psychologists, philosophers, and, of course, computer scientists.

In 1958, F. Rosenblatt, a physiologist, introduced a learning model, called the Perceptron, reflecting the classical neurophysiological understanding of the learning mechanism as an interaction of a large number of simple agents (McCulloch–Pitts model of neuron) with a simple reaction to rewards and punishment as the learning algorithm. The new idea in the Perceptron was its implementation on computers demonstrating that it can generalize.

The Perceptron was not only considered a success in solving one special problem, it was considered a success of an idea of simple organization of a large number of simple agents. The future improvement of learning machines was connected with more accurate analysis of properties of simple agents in the brain and with more accurate analysis of the general rules of their interactions. After the Perceptron there appeared a number of learning models (for example, called Pandemonium and Cortex) where the analysis centered mainly on speculations about the relation of these models to the construction of the brain.

Novikoff's Theorem

Novikoff's theorem gave an alternative approach. According to this theorem, the Perceptron realizes a simple mathematical idea: It maps input vectors into another space in which it constructs a separating hyperplane. Future analysis showed that the generalization ability of Perceptrons can be explained by simple mathematical constructions. This theorem gave the first answer to the question, What is going on?

Novikoff's theorem immediately gave rise to the following questions:

1. If it is important to construct a separating hyperplane in feature space why should this not be done in the most effective way? There are better mathematical methods for constructing separating hyperplanes.
If there exists one separating hyperplane then there exist many of them. Why not choose the optimal one? Control theory demonstrated how much gain can be achieved using optimal solutions.

2. The goal of learning is generalization rather than separating training data. Why does separating training data lead to generalization? Is separating the training data the best way to control generalization?

In other words, from Novikoff's theorem arose the questions whose answers comprise statistical learning theory. The discussion of these questions was one of the main topics of the machine learning seminars.

In the course of these discussions, somebody removed all unnecessary details of the learning problem and concentrated on the core problem—namely, minimizing the risk functional

$$R(\alpha) = \int Q(z, \alpha) dP(z), \quad \alpha \in \Lambda$$

based on empirical data

$$z_1, \dots, z_\ell.$$

Minimization of Risk from Empirical Data and the Classical Statistics Paradigm

Minimization of the risk functional based on empirical data was not considered in great detail in the classical statistical paradigm. Classical tradition is to consider three main statistical problems: density estimation, regression estimation, and discriminant function estimation, separately, using specific parametric models for each of these problems. Of course, three particular settings of the problem, instead of a single general one, were considered not because statisticians did not see this trivial generalization, but because the particular models used in classical statistics for solving the main problems of function estimation based on empirical data did not allow such generalization.

As we mentioned in the Introduction, the classical approach to solving these problems was based on methods developed by R. Fisher in the 1920s and 1930s. Fisher suggested three approaches for solving main function estimating problems based on the maximum likelihood method:

1. He suggested using the maximum likelihood method for estimating parameters a_0 of a density function belonging to a parametric set of densities $p(x, a)$:

$$L(a) = \sum_{i=1}^{\ell} \ln p(x_i, a) \longrightarrow \max_a$$

2. He suggested using the maximum likelihood method for estimating parameters of the regression function, belonging to a parametric set of functions $f(x, a)$. The regression is estimated from data that are values

of the regression function at given points corrupted by additive noise ξ with known density function $p(\xi)$

$$y_i = x_i + \xi_i.$$

Estimating parameters of regression in this case is equivalent to estimating parameters of density $p(\xi) = p(y_i - f(x_i, a))$.

3. He suggested using the maximum likelihood method for estimating parametric densities of different classes $p_k(x, a)$, $k = 1, \dots, m$, in discriminant analysis. The estimated densities are used to construct a discriminant function.

In 1946 Harold Cramer in his famous book *Mathematical Methods of Statistics* (Cramer, 1946), by putting these methods on a firm mathematical basis, created the classical paradigm of applied statistics. The key point in the classical paradigm is analysis of the accuracy of the estimation of the *vector* parameters that specify the unknown functions rather than analysis of performance of the estimated functions. That is why classical statistics did not consider the problem of minimizing the risk functional for a given set of functions.

The problem of estimating functions with good performance rather than parameters of unknown functions became the core problem of statistical learning theory. This problem defined a new development of statistical theory, pushing it toward the theory of function approximation and functional analysis.

Three Elements of Scientific Theory

According to Kant, any theory should contain three elements:

1. Setting of the problem
2. Resolution of the problem
3. Proofs

At first glance, this remark looks obvious. However, it has a deep meaning. The crux of this remark is an idea that these three elements of theory in some sense are independent and equally important.

1. The setting of the problem specifies the models that have to be analyzed. It defines the direction of research.
2. However, the resolution of the problem does not come from deep theoretical analysis of the setting of the problem, but rather precedes this analysis.
3. Proofs are constructed not for searching for the solution of the problem. but for justification of the solution that has already been suggested.

The first two elements of the theory reflect the understanding of the essence of the problem of interest, its philosophy. The proofs make the general (philosophical) model a scientific theory.

Two Resolutions of the Risk Minimization Problem

In Chapter 1, we considered minimization of risk functional on the basis of empirical data as one of two settings of the learning problem (imitation of supervisor). For this setting, we consider the resolution called *the principle of empirical risk minimization*. In order to find the minimum of the expected risk, we minimize the empirical risk functional

$$R_{\text{emp}}(\alpha) = \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha)$$

constructed on the basis of data. In Chapter 6, we made a modification to this resolution: We considered the structural risk minimization principle.

However, these principles (resolutions) are not the only possibilities. An important role in learning processes belongs to the stochastic approximation principle discovered by Robbins and Monroe (1951), where in order to minimize the expected loss functional on the basis of empirical data (in a set of vector-parameterized functions), one uses the following iterative procedure

$$\alpha(n) = \alpha(n-1) - \gamma_n \nabla_{\alpha} Q(x_n, \alpha_{n-1}),$$

where $\nabla_{\alpha} Q(z, a)$ is a gradient (or generalized gradient) of the function $Q(z, a)$ and γ_n is a sequence of constants that depend on n. It was shown that under wide conditions on $Q(z, a)$ and y_n , this procedure converges.

M. A. Aizerman, E. M. Braverman, and L. I. Rozonoer (1965–1967), Amari (1967), and Ya. Z. Tsypkin (1968) constructed the general asymptotic theory of learning processes based on the stochastic approximation induction principle. Later, in 1970–1974 several books were published on this theory (Aizerman, Braverman, and Rozonoer, 1970; Tsypkin, 1971; Tsypkin, 1973).

The stochastic approximation inductive principle, however, cannot be considered as a model for learning from small samples. A more realistic model for these methods is the empirical risk minimization inductive principle. Therefore along with analysis of the stochastic approximation inductive principle the theory of the empirical risk minimization inductive principle had been developed (Vapnik and Chervonenkis, 1968–1974).

The Problem of Density Estimation

The second setting of the learning problem (identification of a supervisor function) is connected with the density estimation problem.

Analyzing the development of a theory of density estimation, one can see

how profound Kant's remark was. Classical density estimation theories, both parametric and nonparametric, contained only two elements: resolution of the problem and proofs. They did not contain the setting of the problem.

In the parametric case, Fisher suggested the maximum likelihood method (resolution of the problem); later it was proved by Le Cam (1953), Ibragimov and Hasminskii (1981), and others that under some (and, as we saw in this book, not very wide) conditions the maximum likelihood method is consistent.

The same happened with nonparametric resolutions of the problem: histogram methods (Rosenblatt, 1956), Parzen's window methods (Parzen, 1962), projection methods (Chentsov, 1963), and so on. First the methods were proposed, followed by proofs of their consistency. In contrast to parametric methods the nonparametric methods are consistent under wide conditions.

The absence of a general setting of the problem made the density estimation methods look like a list of recipes. It also appeared that heuristic efforts make the only possible approach to improve suggested density estimation methods. This created a huge collection of heuristic corrections to nonparametric methods for their practical applications.

The attempt to suggest a general setting of the density estimation problem was done in an article by Vapnik and Stefanuyk (1978) where the density estimation problem was considered, as a problem of solving an integral equation with an unknown right-hand side, but given data. This general setting (which is general because it follows from the definition of density) immediately connected density estimation theory with two fundamental theories:

1. Theory of solving ill-posed problems
2. Glivenko–Cantelli theory

Theory of Ill-Posed Problems

The theory of ill-posed problems can be considered one of the most important achievements in understanding the nature of many problems. Originally it was developed for solving inverse problems of mathematical physics. Later, however, the general nature of this theory was understood. It was demonstrated that one has to take into account the statements of this theory every time when one faces an inverse problem—that is, when one tries to derive the unknown causes from known consequences. In particular, the results of the theory of ill-posed problems are important for statistical inverse problems, one of which is the problem of density estimation.

The existence of ill-posed problems was discovered by Hadamard (1902). Hadamard thought that ill-posed problems are purely mathematical phenomenon and that the real-life problems were well-posed. Soon, however, it was discovered that there exist important real-life problems that are ill-posed.

Tikhonov (1943), proving a lemma about an inverse operator, described the nature of well-posed problems and therefore discovered ways for regularization of ill-posed problems. It took 20 more years before Phillips (1962), Ivanov (1962), Tikhonov (1963), and Lavrentev (1962) came to the same constructive regularization idea described—however, in a slightly different form.

The important message of the regularization theory was the fact that in the problem of solving operator equations

$$Af(t) = F(x),$$

which define an ill-posed problem, the obvious resolution to the problem, namely minimizing the functional

$$R(f) = ||Af - F||^2,$$

does not lead to good results. Instead one should use a nonobvious resolution that suggests minimizing the "corrupted" functional

$$R^*(f) = ||Af - F||^2 + \gamma W(f).$$

These results were the first indication that in function estimation problems obvious resolutions may be not the best.

It should be added that even before the regularization method was introduced, V. Fridman (1956) found regularization properties of stopping early an iterative procedure of solving operator equations (note that here is the same idea: a "corrupted" solution is better than an "uncorrupted" one).

The regularization technique in solving ill-posed problems was not only the first indication of the existence of nonobvious resolutions to the problems that are better than the obvious resolution, but it also gave an idea how to construct these nonobvious resolutions. One can clearly see that many techniques in statistics, and later in learning theory, that construct a better solution to the problem were adopted from the regularization technique for solving ill-posed problems.

Glivenko–Cantelli Theorem and Kolmogorov Bounds

In 1933, the same journal published three articles that can be considered as the cornerstone of statistical science. Glivenko proved that the empirical distribution function always converges to the unknown continuous distribution function, Cantelli proved that the empirical distribution function converges to any unknown distribution function, and Kolmogorov gave an exact asymptotic rate of convergence of the empirical distribution function to the desired continuous distribution function.

The important message of these results is that empirical data contain enough information to estimate an unknown distribution function. As was

Table 1. Structure of the classical theory of statistics and the statistical learning theory

	Classical Statistics Paradigm	Statistical Learning Theory Paradigm
Setting of the problem	Estimation of function parameters	Minimizing expected risk using empirical data
Resolution of the problem	ML method	ERM or SRM methods
Proofs	Effectiveness of parameter estimation	Existence of uniform law of large numbers.

shown in Chapter 2, to estimate the probability measure, one needs to estimate a density function, that is, to solve an ill-posed problem.

This ill-posed problem, however, is not too hard: It is equivalent to estimating the first derivative of a function on the basis of measurements of this function; the measurements are such that when using them one can construct an approximation that converges exponentially fast to the unknown distribution function (for example, this is true for an empirical distribution function). Therefore (according to Chapter 7), using the regularization method for solving an integral equation that defines a density estimation problem, one can construct various methods (classical and new) that estimate density if the latter exists.

Moreover, as we saw in Chapter 2, the solution of the risk minimization problem does not require estimating a probability measure as a whole; it is sufficient to estimate it partially (on some subset of events). The partial estimate defined by a subset as described in the Glivenko–Cantelli theorem is always possible. For this subset, there exist exponential bounds on the rate of convergence, according to the Kolmogorov theorem. One of the goals of learning theory therefore was to obtain the same results for different subsets of events. This goal was achieved almost 40 years after Glivenko–Cantelli–Kolmogorov theorems had been proved.

Therefore, when analyzing the roots of different approaches to a function estimation problem, one can make Table 1, which shows the difference between the classical statistics paradigm and the statistical learning theory paradigm.

CHAPTER 2

In many respects, the foundations of statistical learning theory coincide with the foundations of statistical theory as a whole.

To see this, we have to discuss the foundations of statistics from some

general point of view. In the 1930s, when Glivenko, Cantelli, and Kolmogorov proved their theorems, the most important problem in the foundation of statistics was considered the problem of the nature of randomness. Several points of view were under consideration.

In 1933 Kolmogorov (1933b) introduced the axioms of probability theory. With this axiomatization the probability theory became a purely deductive discipline (say, as geometry) whose development depends only on **formal** inferences from the axioms. This boosted the development of probability theory. However, the theory lost its connection to the physical concepts of randomness—it simply ignored them. Nevertheless, the question "What is randomness?" remained and needed to be answered. Thirty-two years after introducing the axioms, Kolmogorov (1965) suggested the answer: He introduced the algorithmic complexity concept and defined random values as the output of algorithms with high complexity.

Therefore the problem of the foundation of statistics has two natures: (1) the mathematical (formal), connected with **axiomatization** of probability theory, and (2) the physical, describing randomness as an outcome of too complex algorithms.

In Chapter 2, we touched upon the formal (mathematical) part of the problem of the foundation of statistics and its relation to the foundation of learning theory, and in Chapter 6 when we considered the structural risk minimization principle we discussed the relation of the algorithmic complexity concept and the capacity concept.

In Chapter 2, when sketching the main problem of statistics as an estimation of probability measure from a collection of data, we stressed that the foundation of statistics is connected to the problem of estimating the density function in the L_1 norm: If the density function does exist, then there exists the estimator of the probability measure. The connection of estimating the probability measure to density estimation in the L_1 norm was discussed by several authors, including Abou-Jaoude (1976), and Chentsov (1981). In particular, they discussed the conditions for existence of L_1 convergence of the density estimator.

In 1985 Luc Devroye and Laslo Györfi published a book entitled *Nonparametric Density Estimation: The L_1 View*, which presented a comprehensive study of nonparametric density estimation methods. In the presentation in Chapter 2, devoted to convergence of probability measure, I followed the ideas of Chentsov (1988), described in the appendix ("Why the L_1 approach?") to the Russian translation of this book.

Describing a partial estimate of the probability measure over some subset of a sigma algebra, we considered the generalized Glivenko–Cantelli problem. In classical statistics the generalization of this theorem and the corresponding Kolmogorov-type bounds were obtained for multidimensional empirical distribution functions and for a sharp nonasymptotic bound in the one-dimensional case. The main results here were obtained by Dvoretzky, Kiefer, and Wolfowitz (1956) and by Massart (1990). It was shown that the

(nonasymptotic) rate of convergence of empirical distribution function to the actual distribution function has the bound

$$P \left\{ \sup_{x \in R^n} |F(x) - F_{\text{emp}}(x)| > \varepsilon \right\} < 2e^{-2\varepsilon^2\ell}.$$

The general picture of the Glivenko–Cantelli problem as uniform convergence over an arbitrary set of events was established in the 1970s after the main results about uniform convergence of the frequency to their probabilities over a given set of events were obtained and the necessary capacity concepts were introduced.

Therefore, introduction of the generalized Glivenko–Cantelli problem is in some sense a reconstruction of the history. This generalization could have happened before the analysis of the pattern recognition problem was complete, but it did not. It is a direct consequence of the analysis of the ERM principle for the pattern recognition problem.

CHAPTER 3

The presentation of material in Chapter 3 also does not reflect the historical development of the theory. The Key Theorem of learning theory that starts Chapter 3 was proven 20 years after the pattern recognition theory had been constructed (Vapnik and Chervonenkis, 1989). The theory that included the Key Theorem was developed to show that for consistency of the empirical risk minimization induction principle, existence of the uniform law of large numbers is necessary and sufficient; and therefore for any analysis of learning machines that use the ERM principle, one cannot avoid this theory.

The First Results in VC Theory. Late 1960s

Development of statistical learning theory started with a modest result. Recall that in Rosenblatt's Perceptron the feature space was binary (Rosenblatt suggested using McCulloch–Pitts neurons to perform mapping). In the mid-1960s it was known that the number of different separations of the n-dimensional binary cube by hyperplanes is bounded as $N < e^{n^2}$. Using the reasoning of Chapter 4 for the optimistic case in a simple model, we demonstrated (Vapnik and Chervonenkis, 1964) that if one separates (without error) the training data by a hyperplane in n-dimensional binary (feature) space, then with probability $1 - \eta$ one can assert that the probability of test error is bounded as

$$p \leq \frac{\ln N - \ln \eta}{\ell} \leq \frac{n^2 - \ln \eta}{\ell}$$

In this bound the capacity term $\ln N$ has order of magnitude n^2 . Deriving such bounds for nondiscrete feature space, we introduced the capacity con-

cepts used in this book: first the growth function, then the VC dimension, and only after these concepts the entropy of the set of indicator functions. By the end of 1966 the theory of uniform convergence of frequencies to their probabilities was completed. It included the necessary and sufficient conditions of consistency, as well as the nonasymptotic bounds on the rate of convergence.

The results of this theory were published in (Vapnik and Chervonenkis, 1968, 1971). Three years later we published a book (Vapnik and Chervonenkis, 1974) describing the theory of pattern recognition. The 1974 book contained almost all the results on pattern recognition described in the present book.

The generalization of the results obtained for the set of indicator functions to the set of bounded real-valued functions was a purely technical achievement. It did not need construction of new concepts. By the end of the 1970s, this generalization was complete. In 1979 it was published in a book (Vapnik, 1979) that generalized the theory of estimating indicator functions to estimating real-valued functions, and in two years (1981) we published the necessary and sufficient conditions for the uniform law of large numbers (Vapnik and Chervonenkis, 1981). The 1979 book in which the last result was included was translated into English in 1982. It contained almost all the results on real-function estimation described in this book.

Two strong reactions to the mathematical techniques developed appeared in the late 1970s and early 1980s: one at MIT and another at Kiev State University. The fact is that in the theory of probability an important role belongs to the analysis of two problems: the law of large numbers and the central limit theorem. In our analysis of the learning problem we introduced a uniform law of large numbers and described the necessary and sufficient conditions for its existence. The question arose about the existence of the uniform central limit theorem. The discussion about a uniform central limit theorem was started by Dudley (1978).

Using capacity concepts analogous to those developed in the uniform large numbers theory, Kolchinskii (1981), Dudley (1978, 1984), Pollard (1984), and Giné and Zinn (1984) constructed this theory.

Giné and Zinn also extended the necessary and sufficient conditions obtained for the uniform law of large numbers from the sets of uniformly bounded functions to the sets of unbounded functions. These results were presented in Theorem 3.5.

After the discovery of the conditions for the uniform central limit theorem, the uniform analog of the classical structure of probability theory was constructed. For learning theory, however, it was important to show that it is impossible to achieve generalization using the ERM principle if one violates the uniform law of large numbers.

This brought us to the Key Theorem, which points out the necessity of an analysis of uniform one-sided convergence (uniform one-sided law of large numbers). From a conceptual point of view, this part of the analysis was ex-

tremely important. At the end of the 1980s and the beginning of the 1990s there was a common opinion that statistical learning theory provides a pessimistic analysis of the learning processes, the worst-case analysis. The intention was to construct the "real-case analysis" of the ERM principle.

In the Key Theorem of the learning theory it was proven that the theory of the ERM principle which differs from the developed one is impossible. Violation of the uniform law of large numbers brings us to a situation that in the philosophy of science is called a nonfalsifiable theory.

These results brought statistical learning theory to interaction with one of the most remarkable achievements in philosophy in this century: K. Popper's theory of induction.

Now knowing statistical learning theory and rereading Popper's theory of induction, one can see how profound was his intuition: When analyzing the problem of induction without using special (mathematical) models, he discovered that the main concept responsible for generalization ability is the capacity.[†]

Milestones of Learning Theory

At the end of Chapter 3, we introduced three milestones that describe the philosophy of learning theory. We introduced three different capacity concepts that define the conditions (two of them are the necessary and sufficient ones) under which various requirements to generalization are valid. To obtain the new sufficient conditions for consistency of the learning processes, one can construct any measure of capacity on a set of functions that are bounded from below by those defined in the milestone.

Thus, in Chapter 1 we introduced the setting of a learning problem as a problem of estimating a function using empirical data.

For resolution of this problem using the ERM principle, we obtained proofs of consistency using the capacity concept.

CHAPTER 4

The results presented in Chapter 4 mostly outline the results described in Vapnik and Chervonenkis (1974). The only difference is that the constant in the bound (4.25) was improved. In Vapnik and Chervonenkis (1968, 1971, 1974) the bound defined by Theorem 4.1 had constant $c = 1/4$ in front of ε^2 in Eq. (4.46).

In 1991 Leon Bottou showed me how to improve the constant (1 instead

[†] It is amazing how close he was to the concept of VC dimension.

of $1/4$). The bound with improved constant was published by Parrondo and Van den Broeck (1993).

However, for many researchers, the goal was to obtain constant 2 in front of ε^2 as in the asymptotically exact formula given by Kolmogorov. This goal was achieved by Alexander (1984), Devroye (1988), and Talagrand (1994). To get 2 in front of ε^2 , the other components of the bound must be increased. Thus Alexander's bound has too large a constant, and Talagrand's bound contains an undefined constant. In Devroy's bound the right-hand side is proportional to

$$\exp \left\{ \left(\frac{h \left(1 + \ln \frac{\ell^2}{h} \right)}{\ell} - 2\varepsilon^2 \right) \ell \right\}$$

instead of

$$\exp \left\{ \left(\frac{h \left(1 + \ln \frac{2\ell}{h} \right)}{\ell} - \varepsilon^2 \right) \ell \right\}$$

presented in this book (see Eq. (4.46)). Asymptotically, Devroy's bound is sharper. However, for small samples (say, $\ell/h < 20$) the bound given in this book is better for all

$$\varepsilon \leq \sqrt{\frac{\ln \ell / 2}{20}}.$$

Also, the bounds on risk obtained from bounds on uniform convergence with $c = 1$ have clear physical sense: they depend on the ratio ℓ/h .

The important role in the theory of bounds belongs to Theorem 4.3 which describes the structure of the growth function, showing that this function either is equal to $\ell \ln 2$ or can be bounded from above by the function $h(\ln \ell/h + 1)$.

This theorem was published for the first time (without proofs) in 1968 (V. Vapnik and A. Chervonenkis, 1968). Vapnik and Chervonenkis (1971) published the proofs. In 1972, Sauer (1972) and Shelah (1972) independently published this theorem in a form of the combinatorial lemma.

CHAPTER 5

The content of Chapter 5 mostly outlines the results obtained in the late 1970s and published in Vapnik (1979, English translation 1982).

The main problem in obtaining constructive bounds for uniform convergence is generalization of the VC dimension concept for sets of real-valued

functions. There are several ways to make this generalization. In Chapter 5 we used the direct generalization that was suggested in the 1974 book (Vapnik and Chervonenkis, 1974). This generalization led to simple bounds and makes it possible to introduce the bounds on risk for sets of unbounded loss functions.

There exist, however, other ways to generalize the VC dimension concept for sets of real-valued functions. One of these is based on a capacity concept called the VC subgraph, which was introduced by R. Dudley (1978). Using the VC subgraph concept (which was renamed the pseudodimension), Pollard (1984) obtained a bound on the rate of uniform convergence for the set of bounded functions. These results were used by Haussler (1992) to obtain bounds for the rate of generalization of learning machines that implement sets of bounded real-valued functions.

In the distribution free case for sets of indicator functions the finiteness of the VC dimension defines the necessary and sufficient conditions for uniform convergence. For sets of real-valued functions the finiteness of the VC dimension (or the pseudodimension) is only a sufficient condition. The necessary and sufficient conditions are described by a modified version of the VC dimension (or the pseudodimension) (Alon et al. 1993). This modification was suggested by Kearns and Schapire (1994).

CHAPTER 6

The idea of the existence of the advanced induction principle that involved capacity control appeared in the 1960s in different branches of science. First it was introduced by Phillips (1962), Tikhonov (1963), Ivanov (1962), and Lavrentiev (1962) as a method for solving ill-posed problems. Later in the 1970s it appeared in statistics as advanced methods for density estimation: sieve method (Grenander, 1981), penalized method of density estimation (Tapia and Thomson, 1978), and so on. This analysis was done in the framework of asymptotic theory and had described more or less qualitative results.

The quantitative theoretical analysis of the induction principle based on the algorithmic complexity concept was started by Solomonoff (1960), Kolmogorov (1965), and Chaitin (1966). This idea was immediately recognized as a basis for creating a new principle of inductive inference. In 1968 C. Wallace and D. M. Boulton, on the basis of **Solomonoff–Kolmogorov–Chaitin** ideas, introduced the so-called Minimum Message Length (MML) principle (Wallace and Boulton, 1968). Later, in 1978, an analogous principle, called Minimum Description Length, was suggested by Rissanen (1978).

The important result that demonstrated self-sufficiency of the Solomonoff–Kolmogorov–Chaitin concept of algorithmic complexity for induction was obtained by Barron and Cover (1991) for the density estimation problem.

In Chapter 6 we applied the Solomonoff–Kolmogorov–Chaitin ideas to the pattern recognition problem. We showed that the compression idea is self-sufficient in order to obtain the bounds on the generalization ability of the

MML–MDL induction principle for a finite number of observations (Vapnik, 1995).

In Chapter 10 we obtained a bound for SV machines which depends on the minimum of three values, one of which is the number of essential support vectors. Since the ratio of the number of support vectors to the number of observations can be considered as a compression coefficient and the number of essential support vectors is less than the number of support vectors, the obtained bound can be better than the ones that follow from the compression scheme; that is, there is room for searching for an advanced induction principle.

In 1974 using the bound for uniform convergence we introduced the Structural Risk minimization induction principle. In naming this principle we tried to stress the importance of capacity control for generalization. The main difference between the SRM principle and methods considered before was that we tried to control a general capacity factor (e.g., the VC dimension) instead of a specific one (say, the number of parameters). Lugosi and Zegev (1994) proved that SRM principle is universally consistent (see Devroye et al., 1996).

An important feature of the SRM principle is that capacity control can be implemented in many different ways (using different type of structures). This describes the mathematical essence of the SRM principle. The physical essence for this problem is to describe which type of structure is appropriate for our real-world tasks.

When discussing this question, one usually refers to Occam's razor principle:

Entities should not be multiplied beyond necessity.

In other words,

The simplest explanation is the best.

In this book we have encountered several interpretations of the concept of simplest explanation which fit the general SRM scheme. In particular, one can define the concept of *the simplest* as one that (1) has the smallest number of features (free parameters), (2) has the smallest algorithmic complexity, and (3) has the largest margin.

Which of them corresponds to Occam's razor? If we apply Occam's razor principle to the problem of choosing one of these three interpretations, it would choose option 1 (the smallest number of features).

Chapter 12 and especially Chapter 13 demonstrated that algorithms which ignore the number of parameters and control the margin (such as SV machines, neural networks, and AdaBoost schemes) often outperform classical algorithms based on the philosophy of controlling the number of parameters.

In the light of this fact, Occam's razor principle is misleading and perhaps should be discarded in the statistical theory of inference.

The important part of the problem of estimating multidimensional functions is the problem of function approximation. As stated in the Bernstein–Vallee-Poussin theorem, a high rate of approximation can be achieved only for smooth functions. However, the concept of smoothness of functions in high-dimensional spaces (that reflect the same phenomenon) can be described differently. In Chapter 6 we considered the nonclassical smoothness concepts that were introduced by Barron (1993), Breiman (1993), and Jones (1992). For such smooth functions they suggested simple structures with elements described only by the number of basis functions of the type $f((x^* w) + w_0)$.

Another interesting set of functions for which the rate of approximation is fast (except for the logarithmic factor) and the constants in the bound depend on VC dimension of the set of approximating functions was introduced by Girosi (1995).

The idea of local approximation of functions has been considered in statistics for many years. It was introduced in nonparametric statistics as the k -nearest neighbor method for density estimation or as the Nadaraya–Watson method for regression estimation: Nadaraya (1964), Watson (1964). Classical analysis of these methods is asymptotic. In order to achieve the best results in the nonasymptotic case, it is reasonable to consider a more flexible scheme or local model that includes both the choice of the best locality parameter and the complexity of the local model (in the classical consideration, both the locality parameter and complexity of the local model are fixed).

An article by Vapnik and Bottou (1993) reported bounds for such models. Section 6, which is devoted to local learning methods, is based on this article. Now it would be very useful to define the optimal strategy for simultaneous choice of both parameters. This problem is not of just purely theoretical interest, but also of practical importance, since by using local methods one can significantly improve performance. This fact was reported for pattern recognition in an article by Bottou and Vapnik (1992) where, by using a local version of the neural network, performance was improved more than 30%. Using this idea for regression, Bottou, Driancourt, and Ignace constructed in 1994 a system for highway traffic forecast that outperformed the existing system at **Ministère de L'Equipment** (France) by almost 30%.

Remarks on Bayesian Inference

In this book we have not considered the Bayesian approach. However, to better understand the essence of the SRM (or MDL) principle, it is useful to compare it to Bayesian inference.

In the middle of the eighteenth century, Thomas Bayes derived

... the first expression in precise quantitative form of a mode of inductive inference (Encyclopaedia Britannica, 1965).

This expression became one of the most important formulas in probability theory:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

When

$$B = z_1, \dots, z_\ell$$

is a sequence of i.i.d. observation and $P(A)$ is an a priori probability function on a set of vector parameters A that control distribution $P(B|A)$, then

$$P(A|z_1, \dots, z_\ell) = \frac{P(z_1|A) \times \dots \times P(z_\ell|A)P(A)}{P(z_1, \dots, z_\ell)}$$

This formula defines a posteriori probabilities $P(A|z_1, \dots, z_\ell)$ on parameters A (after one takes into account the sequence z_1, \dots, z_ℓ and the a priori probabilities $P(A)$).

The a posteriori probabilities can be used for estimating a function from a given collection of functions. Consider, for simplicity, the problem of regression estimation from a given set of functions $f(x, a), a \in A$, using measurements corrupted by additive noise:

$$y_i = f(x_i, \alpha_0) + \xi_i$$

(here $\alpha_0 \in A$). Using a priori probabilities $P(\alpha)$, a sequence of measurements

$$B = (y_1, x_1), \dots, (y_\ell, x_\ell),$$

and information about the distribution function of the noise $P(\xi)$ one can estimate the a posteriori probability on parameters a that define functions $f(x, a)$ as follows

$$P(\alpha|B) = \frac{\prod_{i=1}^{\ell} P(y_i - f(x_i, \alpha))P(\alpha)}{\int \prod_{i=1}^{\ell} P(y_i - f(x_i, \alpha))P(\alpha) d\alpha}.$$

Let us call ***Bayesian*** any inference that is made on the basis of an a posteriori probability functionⁱ $P(\alpha|B)$.

The Bayesian approach considers the following two strategies of inference from a posteriori probability functions:

ⁱThis is a particular formulation of the Bayesian approach. A more general formulation exists based on concepts of subjective probability and utility.

1. The inference based on the idea of maximization of a posteriori probability (MAP) where one chooses the function that maximizes a posteriori probability $P(\alpha|B)$. This function coincides with the function that maximizes the functional

$$R(\alpha) = \frac{1}{\ell} \sum_{i=1}^{\ell} \ln P(y_i - f(x_i, \alpha)) + \frac{1}{\ell} \ln P(\alpha). \quad (\text{a})$$

2. The inference based on the idea of averaging an admissible set of functions over a posteriori probability where one constructs an approximating function that is the average of functions from an admissible set with respect to a posteriori probabilities:

$$\phi(x|B) = \int f(x, \alpha) P(\alpha|B) d\alpha. \quad (\text{b})$$

The constructed function has the following remarkable property: it minimizes the expectation of quadratic loss from admissible regression functions where the expectation is taken both over the training data and over the set of admissible functions:

$$\Phi(\phi) = \int (f(x, \alpha) - \phi(x|B))^2 P(B|\alpha) P(\alpha) dx d\alpha dB. \quad (\text{c})$$

The fact that $\phi(x|B)$ minimizes the functional $\Phi(\phi)$ is often considered as the justification of the averaging Bayesian inference.

In this book we consider consistency as the smallest requirement for function estimation methods.

One can show that the method of maximum of a posteriori probability is always consistent if the set of admissible functions contains a countable number of elements. Compare this to the maximum likelihood method (minimum empirical risk method) that is always consistent only if the set of admissible functions is finite. However, for the uncountable infinite set of admissible functions the method of maximum a posteriori probability is not necessarily consistent. To construct a consistent method, one should modify the MAP method in the spirit of SRM in order to control capacity. (Recall that the SRM method is strongly universally consistent.)

In spite of the fact that the averaging method minimizes functional (c), it is not necessarily consistent as well. See the article by Diaconis and Freedman (1986) for details. This means that to make the averaging method universally consistent, one has to modify it.

More important, however, is the fact that to use Bayesian inferences, one must have strong a priori information:

1. The given set of functions of the learning machine coincides with a set of target functions (qualitative a priori information).

2. A priori probabilities on the set of target functions are described by the given expression $P(\alpha)$ (quantitative a priori information).

Statement 2 (quantitative information) is not as important as statement 1 (qualitative information). One can prove that, under certain conditions, if quantitative a priori information is incorrect (but qualitative is correct), then with an increasing number of observations the influence of incorrect information decreases. Therefore, if the qualitative information is correct, the Bayesian approach is appropriate.

The requirement for the correctness of the qualitative information is crucial: The set of target functions **must** coincide with the set of functions of a learning machine. Otherwise Bayes' formula has no sense.

Therefore in the framework of Bayesian inference one cannot consider the following problem: Find the function that approximates a desired one in the best possible way if a set of functions of a learning machine does not coincide with a set of target functions. In this situation, any chosen function from the admissible set has zero a priori probability; consequently, according to Bayes formula, the a posteriori probability is also equal to zero. In this case inference based on a posteriori probability function is impossible.

In contrast to the MAP method, the capacity (algorithmic complexity) control methods SRM (or MDL) use weak a priori information about reality: They use a structure on the set of functions (the set of functions is ordered according to the idea of usefulness of the functions) and choose the appropriate element of structures by capacity control. To construct a structure on the set of functions, one does not need to use information that includes an exact description of reality. Therefore, when using the SRM (MDL) approach, one can approximate a set of functions that is different from the admissible set of functions of the learning machines (the appropriate structure affects the rate of convergence; however, for any admissible structure the SRM method is consistent).

The difference between the SRM approach and the MAP approach is in the following: The SRM approach does not need a priori information about the set of target functions due to capacity control, while the MAP approach does not include the capacity control but uses such a priori information. The capacity control makes the SRM method universally consistent, while even the correct a priori information about the set of target functions does not guarantee consistency of the MAP method.

The averaging method has a more general framework than averaging (b) with respect to the posteriori probability. One can consider this method as constructing a hyperplane in Hilbert space, where using training data B one defines a function $P(\alpha|B)$ that specifies the hyperplane. The specific feature of the averaging method is that function $P(\alpha|B)$ has to be non-negative.

In the Bayesian approach, one estimates this function with posteriori probabilities, which limits this type of inference due to the necessity of using the

Bayes inversion formula. It is possible however to construct averaging methods that do not rely on the Bayesian formula (see V. Vovk, 1991).

In learning theory several ideas of averaging were introduced that also do not rely on averaging according to posteriori probabilities, including the Bagging averaging suggested by L. Breiman (1996) and the AdaBoost averaging suggested by Y. Freund and R. Schapire (1995).

As in the SVM theory for pattern recognition the generalization ability of averaging methods was shown also to be controlled by the value of the margin.

Suppose we are given training data

$$(y_1, x_1), \dots, (y_\ell, x_\ell), \quad y_i \in \{-1, 1\}$$

and a set of n indicator functions $u_1(x), \dots, u_n(x)$.

Consider the n dimensional vector (feature vector)

$$\mathbf{u}_i = (u_1(x_i), \dots, u_n(x_i))$$

and the training data in the corresponding feature space

$$(y_1, \mathbf{u}_1), \dots, (y_\ell, \mathbf{u}_\ell)$$

It was shown (Shapire et al., 1997) that AdaBoost constructs the averaging rule

$$f(x) = \sum_{k=1}^n \beta_k u_k(x), \quad \beta_k \geq 0,$$

that separates the training data in the feature space

$$y_i(\mathbf{u}_i * \mathbf{w}) \geq 1, \quad \mathbf{w} \geq 0$$

and minimizes L_1 norm of nonnegative vector $\mathbf{w} = (w_1, \dots, w_n)$

$$\sum_{k=1}^n w_k, \quad w_k \geq 0,$$

In the more general case one can suggest minimizing the functional

$$\sum_{k=1}^n w_k + C \sum_{i=1}^\ell \xi_i, \quad w_k \geq 0$$

subject to constraints

$$y_i(\mathbf{u}_i * \mathbf{w}) > 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \dots, \ell. \quad (\text{d})$$

To construct the averaging hyperplane one can also use the L_2 norm for the target functional (the SVM type approach). In this case one has to minimize the functional

$$(w * w) + C \sum_{i=1}^{\ell} \xi_i, \quad w_k \geq 0$$

subject to constraints (d), where vector w must be nonnegative.

The solution to this problem is the function

$$f(x) = (w * u(x)) = \sum_{i=1}^{\ell} \alpha_i (u(x)^* u_i) + (\Gamma^* u(x)),$$

where the coefficients $\alpha_i, i = 1, \dots, \ell$ and the vector $\Gamma = (\gamma_1, \dots, \gamma_n)$ are the solution of the following optimization problem:

Maximize the functional

$$\begin{aligned} W(\alpha, \Gamma) = & \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j (u_i * u_j) \\ & - \sum_{i=1}^{\ell} \alpha_i (u_i^* \Gamma) - \frac{1}{2} (\Gamma^* \Gamma) \end{aligned}$$

subject to constraints

$$\begin{aligned} 0 \leq \alpha_i \leq C, \quad i = 1, \dots, \ell \\ \gamma_j \geq 0, \quad j = 1, \dots, n \end{aligned}$$

The important difference between the optimal separating hyperplanes and the averaging hyperplanes is that the averaging hyperplanes must satisfy one additional constraint, the coordinates of the vector w must be nonnegative. This constraint creates a new situation both in analysis of the quality of constructed hyperplanes and in optimization techniques. On one hand it reduces the capacity of linear machines which can be used for effective capacity control. On the other hand the existing techniques allow us to average over a small amount of decision rules (say, up to several thousand).

The challenge is to develop methods that will allow us to average over large (even infinite) numbers of decision rules using margin control. In other words, the problem is to develop efficient methods for constructing averaging hyperplanes in high dimensional spaces, analogous to the SVM methods, where there are no constraints on the coefficients of hyperplanes.

CHAPTER 7

The generalization of the theory of solving ill-posed problems originally introduced for the deterministic case to solve stochastic ill-posed problems

is very straightforward. Using the same regularization techniques that were suggested for solving deterministic ill-posed problems and also using the same key arguments based on the lemma about the inverse operator, we generalized the main theorems about the regularization method (Vapnik and Stefanuyk, 1976) to a stochastic model. Later, Stefanuyk (1986) generalized this result for the case of an approximately defined operator.

It was well known that the main problem of statistics—estimating a density function from a more-or-less wide set of functions is ill-posed. Nevertheless the analysis of methods for solving density estimation problem was not considered from the formal point of view of regularization theory.

Instead, in the tradition of statistics one first suggests some method for solving this problem, then proves its **favorable** properties, and then introduces some heuristic corrections to make this method useful for practical tasks (especially for multidimensional problems).

The attempts to derive new estimators from a more general point of view of solving the stochastic ill-posed problem was started with analysis of the various algorithms for density estimation (Aidu and Vapnik, 1989). It was observed that almost all classical algorithms (such as Parzen windows, projection methods) can be obtained on the basis of the standard regularization method of solving stochastic ill-posed problems under conditions that one chooses the empirical distribution function (which is a discontinuous function) as an approximation to an unknown distribution function (which is a continuous function). In Chapter 7 we constructed new estimators using the continuous approximation to the unknown distribution function.

The real challenge, however, is to find a good estimator for multidimensional density defined on bounded support. To solve this problem using ideas described in Chapter 7, one has to define a good approximation to an unknown distribution function: a continuous monotonic function that converges to the desired function with an increasing number of observations as fast as an empirical distribution function converges. Note that for a fixed number of observations the higher the dimensionality of space, the "less smooth" the empirical distribution function. Therefore in the multidimensional case using smooth approximations to the unknown **continuous** distribution function is very important.

On the other hand, for a **fixed** number of observations, the larger the dimensionality of the input space, the greater the number of observations that are "border points." This makes it more difficult to use the smoothness properties of distribution functions. That is, it is very hard to estimate even smooth densities in more or less high-dimensional spaces.

Fortunately, in practice one usually needs to know ***the conditional density*** rather than the density function. One of the ideas presented in Chapter 7 is the estimation of the conditional density (conditional probability) function without estimating densities. The intuition behind this idea is the following: In many applications the conditional density function can be approximated well in low-dimensional space even if the density function is a high-dimensional

function. Therefore the density estimation problem can be more complicated than the one that we have to solve.

The low-dimensional approximation of conditional density is based on two ideas (Vapnik, 1988):

1. One can approximate conditional density locally (along a given line passing through the point of interest).
2. One can find the line passing through the point of interest along which the conditional density (or conditional probability) function changes the most.

Note that if the regression function (or the Bayesian decision rule) can be approximated well by a linear function, then the desired direction is orthogonal to the linear approximation. Therefore one can split the space into two subspaces: One of these is defined by the direction of the approximated linear function, and the other is an orthogonal complement to the first one.

The idea using linear regression is not very restrictive because when using the SV machine, one can perform this splitting into feature spaces.

CHAPTER 8

Transductive inference was discussed for the first time in my 1971 booklet devoted to the problem of pattern recognition. Since that time the discussion on transductive inference was repeated in our 1974 book (Vapnik and Chervonenkis, 1974) and in my 1979 book (Vapnik, 1979) almost without any significant modifications (for English translation of my 1979 book I added sections devoted to estimating real values at given points). Chapter 8 repeats the content of the corresponding chapter from the English edition of the 1979 book.

In spite of the fact that transductive inference can be considered as one of the most important directions of development of statistical learning theory, which should have a strong influence not only on technical discussions on methods of generalization but also on the understanding of ways of inference of human beings, there was almost nothing done in the development of this inference. There is only one article by Vapnik and Sterin (1977) applying transductive inference. In this article by using the transductive version of generalized portrait algorithms (the linear SV machine in input space), the advantage of this type of inference against inductive inference for small sample size was demonstrated: For some real-life pattern recognition problems, the number of test errors was significantly reduced.

In 1976 the generalized portrait algorithm was restricted to constructing an optimal hyperplane in the input space. Now by using generalization of this algorithm, the SV method, one can develop general types of transductive algorithms.

Speculations on Transductive Inference: Inference Through Contradiction

Let me suggest a speculation on transductive inference which I believe reflects nonstandard ways of human inference.

Suppose we are given simultaneously the training set

$$(\omega_1, x_1), \dots, (\omega_\ell, x_\ell),$$

the test set

$$x_{\ell+1}, \dots, x_{\ell+k},$$

and the admissible set of indicator functions $f(x, a)$, $a \in A$.

The discrete space $x_1, \dots, x_{\ell+k}$ factorizes our set of functions into a finite number of equivalence classes F_1, \dots, F_N ($F_k = \{f(x, a) : a \in \Lambda_k\}$).

For solving the problem of estimating the values of functions at the given points, let us consider the empirical risk minimization principle[†]: among N equivalence classes we look for the decision class that separates the training set with zero errors. Let F_1, \dots, F_n be such equivalence classes. If $n > 1$ we have to choose one class among these n equivalence classes.

In Chapter 8, along with the special concept of power equivalence classes for a linear set of functions, we describe a maximum a posteriori probability (MAP)-type approach as a way of introducing a "smart" concept of power of equivalence classes.

Let us repeat it once more. Suppose that there exists a generator which randomly and independently generates $\ell + k$ vectors constructing the current discrete space $x_1, \dots, x_{\ell+k}$.

Suppose also that there exist a priori distributions $P(a)$ on the set of functions $\omega = f(x, a)$, $a \in A$ which determine the desired classifier.

We considered the following scheme:

1. First, a random current discrete space appears.
2. Then this discrete space is randomly divided into two subsets: a subset that contains ℓ vectors and a subset that contains k vectors.
3. According to the distribution $P(a)$, the desired classification function $f(x, a)$ is chosen. Using this function, one determines the values of functions at the points of the first subset, which forms the training set.

The problem is to formulate a rule for estimating the values of the desired function on the second subset which guarantees the minimal expected number of errors.

The MAP solution of this problem is as follows:

[†] We consider here the empirical risk minimization principle only for simplicity. It is more interesting to use the structural risk minimization principle.

1. Consider as the power of the equivalence class the a priori distribution of probability that the desired function belongs to a given equivalence class

$$\text{Power}(F_k) = \int_{\Lambda_k} dP(\alpha), \quad k = 1, 2, \dots, N;$$

here a *priori* means *after* defining the current discrete space, but *before* splitting it on the training and the test sets.

2. Choose among the equivalence classes one which separates the training set without error and has the largest power.

Although the formal solution is obtained, the MAP approach actually does not solve the conceptual problem—namely, to introduce an appropriate concept of power of equivalence classes. It only removes the decision step from the formal part of the problem to its informal part. Now everything depends on the a priori information—that is, the distribution function $P(\alpha)$. The specification of this function is considered as a problem that should be solved outside the MAP inference. Thus, to solve the problem of transductive inference in the framework of the MAP approach, one has to develop a general way for obtaining a priori information.

Of course, nobody can generate a priori information. One can only transform the information from one form to another. Let us consider the following idea for extracting the needed information from the data.

Assume that we have the a priori information in the following form. We are given a (finite) set of vectors

$$x_1^*, \dots, x_r^*$$

(let us call it the universe) which does not coincide with the discrete space but is in some sense close to any possible current space.

For example, if we try to solve the problem of digit recognition, the universe can be a set of signs with approximately the same topology and the same complexity. In other words, the universe should contain the examples which are not digits but which are made in the spirit of digits—that is, have approximately the same features, the same idea of writing, and so on. The universe should reflect (in examples) our general knowledge about real life, where our problem can appear,

We say that the vector x_i^* from our universe is contradictory to the equivalence class F_j if in this class there exists an indicator function that takes 1 on vector x_i^* and there also exists a function that takes value 0 on this point.

We define the power of the equivalence class as the number of examples from the universe which is contradictory to this equivalence class.

Now we can describe the hypothetical transduction step in the following way.

Try to find among the equivalence classes separating the training set one which has the maximum number of contradictions in the universe.[†]

Thus, we split the definition of the power of equivalence classes into two parts: informal and formal. The informal part is the content of our universe and the formal part is the evaluation of the power of equivalence classes on the basis on the universe.

The idea considering the contradictions on the universe as a measure of power of equivalence classes goes in the same direction as (a) **measure** of power for equivalence classes that give solutions for a linear set of functions (based on the margin; see Section 8.5, Chapter 8) or (b) a priori probability of an equivalence class in the MAP statement.

The idea of such an inference can be described as follows:

Be more specific; try to use a solution which is valid for current discrete space and which does not have a sense out of current space.

Or it can be described in the spirit of Popper's ideas as follows:

Try to find the most falsifiable equivalence class which solves the problem.

Of course, from the formal point of view there is no way to find how to choose the universe, as well as no way to find the a priori distribution in the MAP scheme.

However, there is a big difference between the problem of constructing an appropriate universe and the problem of constructing an appropriate a priori distribution in the MAP scheme.

The universe is knowledge about. ***an admissible collection of examples***, whereas the a priori distribution is knowledge about an admissible set of decision functions.

People probably have some feeling about a set of admissible examples, but they know nothing about a distribution on admissible decision functions.

CHAPTER 9

In 1962 Novikoff proved a theorem that bounds the number of corrections of the Perceptron

$$M \leq \frac{D^2}{\rho^2}$$

[†]Of course, this is only a basic idea. Deep resolution of the problem should consider the trade-offs between the number of errors on training data and the number of contradiction examples of the universe. This trade-off is similar to those for induction inference.

Using this bound we showed that in the on-line learning regime perceptron can construct a separating hyperplane with an error rate proportional to $M \ln M / \ell$.

How good is this bound?

Let us consider a unit cube of dimensionality n and all hyperplanes separating vertices of this cube. One can show that it is possible to split the vertices into two subsets such that the distance between corresponding convex hulls (margin) is proportional to 2^{-n} . Therefore in the general case, the a priori bound on quality of the hyperplane constructed by a perceptron is very bad. However for special cases when the margin p is large (the number of corrections M is small), the bound can be better than the bounds that depend on dimensionality of space n . The mapping of input vectors into feature space can be used to create such cases.

Therefore the very first theorem of learning theory introduced a concept of margin that later was used for creating machines with linear decision rules in Hilbert spaces.

However, in the framework of the theory of perceptrons the idea of controlling the margin was not considered. The analysis of perceptrons mainly concentrated on the fact that sets of decision rules of the perceptron are not rich enough to solve many real-life problems. Recall that Rosenblatt proposed to map input vectors into a binary feature space with a reasonable number of features.

To take advantage of a rich set of functions, one can either increase the dimensionality of the feature space (not necessarily binary) and control the generalization using both the value of margin and the value of empirical risk (this idea was later realized in the support vector machines), or construct a multilayer perceptron (neural network) with many controlled neurons.

In 1986 in two different publications, Rumelhart, Hinton and Williams (1986) and LeCun (1986), the back-propagation method of training multilayer networks was introduced. Later it was discovered that Bryson et al. (1963) had described the back-propagation algorithm with Lagrange formalism. Although their description was in the framework of optimal control (they considered a multistage system defined as a cascade of elementary systems) the resulting procedure was identical to back-propagation.

Discovering the back-propagation algorithm made the problem of learning very popular. During the next 10 years, scores of books and articles devoted to this subject were published. However, in spite of the high interest of the scientific community with regard to neural networks, the theoretical analysis of this learning machine did not add much to the understanding of the reason of generalization. Neural network technology remains an art in solving real-life problems.

Therefore at the end of the 1980s and the beginning of the 1990s researchers started looking for alternatives to back-propagation neural networks. In particular, the subject of special interest became the Radial Basis Function method. As we have discussed already in the main part of the book,

the idea of radial basis functions can be clearly seen in the method of potential functions introduced in 1965 (Aizerman, Braverman, and Rozonoer, 1965). The analysis of this method concentrated on on-line learning procedures, while the analysis of radial basis functions was done in the framework of off-line learning. See Powell (1992) for details.

CHAPTER 10

As soon as experiments with the Perceptron became widely known, the discussion on improvement of the Perceptron algorithm started. In the beginning of the 1960s there were many iterative, mostly on-line, methods that later were summarized in books by Tsyplkin (Tsyplkin, 1971, 1973) as a realization of the idea of stochastic approximation. At the same time, the off-line methods for constructing hyperplanes were also under investigation. In 1963 the method of Generalized Portrait for constructing the optimal separating hyperplane in dual space was suggested (Vapnik and Lerner, 1963, Vapnik and Chervonenkis, 1964). This method actually is the support vector method for constructing an optimal hyperplane in the separable case considered in Section 10.1 of Chapter 10. In many practical applications we saw the advantage of an optimal hyperplane compared to a nonoptimal separating hyperplane. In our 1974 book (Vapnik and Chervonenkis, 1974) we published Theorem 10.7, according to which, the generalization ability of the optimal hyperplane that separates training data without errors depends on the expectation of the random variable $\min(\mathcal{D}^2/\rho^2, N, n)$.

In 1974 this theorem (without kernel technique) had limited applications: It could be applied only to the linearly separable pattern recognition problem. Nevertheless, it showed that for this case, classical understanding of the reason for generalization, which relies on the ratio of the number of parameters to the number of observations, does not contain all the truth. There are other factors. The problem was whether these factors reflect the nature of the generalization problem or whether they reflect pure mathematical artifacts.

The SV method demonstrated that these factors must be considered as factors that control generalization ability. To use these factors more efficiently we map relatively low-dimensional input vectors into a high-dimensional feature space where we construct the optimal separating hyperplane. In this space we ignore the dimensionality factor and rely on two others (while the classical approach ignored two other factors).

To make this idea practical we use kernel representation of the inner product based on Mercer's theorem. Before commenting on this idea, let me make one remark.

In Chapter 12 we described experiments, where in order to achieve high generalization we constructed a polynomial of degree 9 in 400-dimensional input space. That is, in order to achieve good performance we separated

712 COMMENTS AND BIBLIOGRAPHICAL REMARKS

60,000 examples in a 10^{23} dimensional feature space. The good generalization was obtained due to the optimality of the constructed hyperplane.

Note that the idea of constructing separating polynomials was discussed in the pattern recognition methods since Fisher suggested discriminant analysis. The main idea of discriminant analysis is the following: Given two normal laws

$$N(\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{n/2} \sqrt{|\Sigma_i|}} \exp \left\{ -\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) \right\}, \quad i = 1, 2$$

that describe the distribution of two classes of instances and given probability p of occurrence of instances of the first class ($q = 1 - p$ for the second), construct the best (Bayesian) decision rule.

The optimal decision rule for this problem is the following quadratic discriminant function:

$$F_2(x) = \theta \left\{ \frac{1}{2}(x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1) - \frac{1}{2}(x - \mu_2)^T \Sigma_1^{-1} (x - \mu_2) + \frac{1}{2} \ln \frac{|\Sigma_1|}{|\Sigma_2|} - \ln \frac{p}{q} \right\}$$

In the particular case when $\Sigma_1 = \Sigma_2 = \Sigma$, the quadratic discriminant function reduces to a linear one:

$$F_1(x) = \theta \left\{ (\mu_2 - \mu_1)^T \Sigma^{-1} x - \frac{1}{2} (\mu_1^T \Sigma^{-1} \mu_1 - \mu_2^T \Sigma^{-1} \mu_2) - \ln \frac{p}{q} \right\}.$$

The idea of the classical discriminant method is to estimate the parameters of the distribution function $\mu_1, \mu_2, \Sigma_1, \Sigma_2, p$ and then put them into an expression for the optimal discriminant function (the so-called substitution method). Of course it is not very difficult to prove that when the number of observations is sufficiently large, this method will give good results.

Fisher, however, understood that for practical reasons the sufficient number of observations has to be large and suggested that we use the linear discriminant rule even if $\Sigma_1 \neq \Sigma_2$. He proposed to construct the artificial covariance matrix $\Sigma = y\Sigma_1 + (1 - y)\Sigma_2$ and substitute it into an expression for the linear discriminant rule. Anderson and Bahadur (1966) solved the problem of choosing a coefficient y that defines the optimal decision rule among linear rules in the case when the best (Bayesian) rule is quadratic. When the dimensionality of the input space exceeds several dozens, the linear discriminant function is used.

Therefore from the very beginning of discriminant analysis, Fisher understood the overfitting problem; and even in the case when the optimal decision rule is quadratic, he preferred a linear discriminant function. The SV method can ignore Fisher's concern, due to optimality of the hyperplane in the corresponding feature space.

To construct a hyperplane in high-dimensional feature space, we use a general representation of the inner product in Hilbert spaces. According to Mercer's theorem, an inner product in Hilbert spaces has an equivalent representation in kernel form. This fact was established by Mercer in 1909 (Mercer, 1909). Since then the Mercer theorem, the related theory of positive definite functions, and the theory of reproducing kernels Hilbert spaces have become important topics of research [see Aronszajn (1943), Steward (1976), Mitchelli (1986), Wahba (1990)]. In particular, this theorem was used to prove the equivalence between the method of potential functions and Rosenblatt's perceptron (Aizerman, Braverman, and Rozonoer, 1964).

Therefore by the mid-1960s, two main elements of the SV machine (the expansion of the optimal hyperplane on support vectors and the constructing hyperplane in feature space using Mercer kernels) were known. It needed only one step to combine these two elements. This step, however, was done almost 30 years later in an article by Boser, Guyon, and Vapnik (1992).

After combining the SV expansion with kernel representation of the inner product, the main idea of the SV machine was realized: One could construct linear indicator functions in high-dimensional space that had a low capacity. However, one could construct these hyperplanes (or corresponding kernel representation in input space) only for the separable case.

The extension of the SV technique for nonseparable cases was obtained in an article by Cortes and Vapnik (1995).

After the SV technique was discovered, the generalization ability of some other learning techniques also was explained by the margin concept rather than by the number of free parameters. Bartlett (1997) proved this fact for neural networks, and Schapire, Freund, Bartlett, and Lee (1997) proved it for the so-called AdaBoost learning technique.

This technique was used by Scholkopf, Smola, and Müller (1997) for constructing nonlinear component analysis by providing linear component analysis in feature space.

Remark

It should be noted that Theorem 10.6 gives a hint that more advanced models of generalization may exist than one based on maximization of the margin. The error bound for optimal hyperplanes described in Theorem 10.6 depends on the expectation of the ratio of two random variables: the diameter of the sphere that contains the support vectors to the margin.

It is quite possible that by minimizing this ratio one can control the generalization better than by maximizing the margin (the denominator of the ratio).

Note that in a high dimensional feature space, where a SV machine constructs hyperplanes, the training set is very sparse and therefore the solution that minimizes this ratio can be very different from one that maximizes the margin.

CHAPTER 11

The generalization of SV machines for estimating real-valued functions was done in the book *The Nature of Statistical Learning Theory* (Vapnik, 1995). It contained a new idea, namely, the ϵ -insensitive loss function. There were two reasons for introducing this loss function:

- (1) to introduce a margin in order to control the generalization of the learning machine and
- (2) to trade the accuracy of approximation for simplicity of solution (for the number of SVs).

With this generalization the SV method became a general method for function representation in high-dimensional spaces which can be used for various problems of function estimation including problems of density estimation and solving linear operator equations [Vapnik, Golowich, and Smola (1997)].

For solving ill-posed problems there exists one more reason to use the ϵ -insensitive loss function: By choosing different values of ϵ for different points of observation, one can better control the regularization process.

CHAPTER 12

As we mentioned, in 1986 the back-propagation method for estimating parameters of multilayer Perceptrons (neural networks) was proposed. In spite of the fact that this had almost no impact on the theory of induction or understanding of the reasons for generalization (if one does not take into account speculations about imitation of the brain), the discovery of neural networks should be considered as a turning point in the technology of statistical inference.

In many examples it was demonstrated that for real-life (rather than small artificial) problems, neural networks give solutions that outperform classical statistical methods. The important insight was that to obtain good performance it was necessary to construct neural networks that contradicted an existing paradigm in statistics: The number of parameters of well-performing neural networks was much larger than would be expected from classical statistics recommendations. In these networks, to obtain good generalization, one incorporated some heuristics both in the architecture (construction) of the network and in the details of the algorithms. Ignoring these heuristics decreased the performance.

Therefore many researchers consider neural network applications to real-life problems to be more art than science.

In 1985 Larry Jackel headed the Adaptive System Research Department at AT&T Bell Laboratories which Yann LeCun joined in 1989. Since that time, the department has become one of the most advanced centers in the

art of solving one specific real-life problem using neural networks, namely, the problem of handwritten digit recognition.

To achieve the highest level of performance for this task, a series of neural networks called LeNet were constructed starting from LeNet 1 (1989), a five-layer convolution neural network, up to seven-layer LeNet 5 (1997) in which, along with classical neural network architecture, various elements of learning techniques were incorporated (including capacity control, constructing new examples using various distortions and noise models). See LeCun et al. (1998) for detail.

Many years of racing for the best results in one specific application is extremely important, since starting from some level any significant improvement in performance can be achieved only due to the discovery of new general techniques. The following accomplishments based on new general techniques have been obtained by this group during the past eight years.

For a relatively small (7,000 of training data) postal service database, the results and corresponding techniques are as follows:

1. Convolutional network (1989)—5.1% error rate
2. Local learning network (1992)—3.3% error rate
3. Tangent distance in the nearest-neighbor method (1993)—2.7% error rate.

The last performance is close to the level of human performance for this database.

Since 1994, experiments also have been conducted using a relatively large (60,000 training examples) NIST database, where the following accomplishments were achieved:

1. Convolutional network (LeNet 1)—1.7%
2. Convolutional network with controlled capacity (LeNet 4)—1.1%
3. Boosted scheme of three networks LeNet 4 with controlled capacity—0.7%
4. Network providing linear transformation invariant LeNet 5—0.9%
5. Network providing linear transformation invariants and elements of noise model LeNet 5a—0.8%.

These results were also on the level of human performance for the NIST database.

Therefore it was challenging to compare the performance of the SV machines with the results obtained using machines of LeNet series. In all experiments with the SV machine reported in Chapter 12 we used a standard polynomial machine where we chose an appropriate order of polynomial and parameter C defining box constraints in the quadratic optimization problem.

The experiments described with SV machines were started by B. Boser and I. Guyon and were then continued by C. Cortes, C. Burges, and B. Scholkopf.

Using standard SV polynomial machines, C. Burges and B. Scholkopf (1997) and Scholkopf et al. (1998) achieved results that are very close to the best for the databases considered.

1. 2.9% of error rate for postal service database (the record is 2.7% obtained using tangent distance)
2. 0.8% of error rate for NIST dataset (the record is 0.70% obtained using three LeNet 4 machines combined in a boosted scheme and using additional training data generated by tangent vectors (Lie derivatives)).

In both cases the best result was obtained due to incorporating a priori information about invariants. For the SV machine, this fact stresses the importance of the problem of constructing kernels that reflect a priori information and keep desired invariants. In experiments with SV machines, however, we did not use this opportunity in full measure.

The important point for the digit recognition problem (and not only for this problem) was the discovery of the tangent distance measure by Simard et al. (1993). The experiments with digit recognition gave the clear message that in order to obtain a really good performance it is necessary to incorporate a priori information about existing invariants. Since this discovery, the methods for incorporating invariants were present in different forms (using boosting procedure, using virtual examples, constructing special kernels) in all algorithms that achieved records for this problem.

CHAPTER 13

Estimation of a real-valued function from a given collection of data traditionally was considered the central problem in applied statistics. The main techniques for solving this problem, the least-squares method and least modulus method, were suggested a long time ago by Gauss and Laplace. However, its analysis started only in our century. The main results in justification of these methods, however, were not unconditional. All theorems about favorable properties of these methods contain some restrictive conditions under which the methods would be considered optimal. For example, theorems that analyze the least-squares method for estimating linear functions state the following:

Among all linear and unbiased methods of estimating parameters of linear functions the least-squares method has the smallest variance (Markov–Gauss theorem). However, why does the estimator have to be linear and unbiased?

If one estimates parameters of the linear function corrupted by additive normal noise, then among all unbiased estimators the least-squares estimator has the smallest variance. Again, why should the estimator be unbiased? And why should the noise be normal?

Under some conditions on additive noise the least-squares method provides the asymptotic unbiased estimate of parameters of linear functions that has the smallest variance. Which method is the best for a fixed number of observations?

James and Stein (1961) constructed an estimator of the mean of random ($n \geq 3$)-dimensional vectors distributed according to the normal law with unit covariance matrix that was biased and that for any fixed number of observations is uniformly better than the estimate by the sample mean. (This result is equivalent to the existence of a biased estimator that is uniformly better than a linear estimator obtained by the least-squares method in the normal regression problem). Later, Baranchik introduced a set of such estimators that contained James–Stein's estimator. The main message from these analyses was that to estimate linear regression well one needs to consider a biased estimator.

In the 1960s the theory of solving ill-posed problems suggested specific methods for constructing biased estimators by using regularization terms. Later in the framework of statistical learning theory, the idea of regularization was used for regression estimation problems.

Classical statistics concentrated on the problem of model selection where to find an estimate of the linear in its parameters function one has to first specify appropriate basis functions (including the number of these functions) and then to estimate a linear function in each basis function. This method, generally speaking, also constructed a biased estimator of parameters.

Note that none of these approaches did develop an exact method (with all parameters fixed) for controlling the desired value of bias. Instead, semi-theoretical approaches were developed for selecting a regularization parameter (in solving ill-posed problems) and choosing appropriate elements of a structure (in the SRM method).

Therefore it is extremely important to compare them experimentally. The results of this comparison (which was done by Cherkassky and Mulier (1998)) are presented in Chapter 13.

Later, Cherkassky and Mulier (1998) demonstrated that using this bound for choosing an appropriate number of wavelets in the wavelet decomposition of the signals outperforms the technique specially constructed for this purpose.

The most interesting part of capacity control is when the capacity of the structure differs from the number of free parameters, for example, defined by the regularization term as in methods for solving an ill-posed problem.

For this situation, it was suggested using the same formulas where instead of the number of parameters, one used the so-called "effective num-

ber of parameters." In formulas obtained from statistical learning theory, capacity is defined by the VC dimension that does not necessarily coincide with the number of parameters.

It is not easy, however, to obtain the exact estimate of the VC dimension for a given set of functions. In this situation the only solution is to measure the VC dimension of the sets of elements of the structure in experiments with a learning machine. The idea of such experiments is simple: The deviation of the expectation of the minimum of empirical risk from 1/2 for training data with random (with probability 1/2) choice of labels depends on the capacity of the set of functions. The larger the capacity, the larger the deviation. Assuming that for maximal deviation there exists a relation of equality type (for small samples, the theory can guarantee only a relation of an inequality type) and that as in the obtained bound, the expected deviation depends on one parameter, namely ℓ/h , one can estimate the universal curve from a machine with known VC dimension and then use this curve to estimate the VC dimension of any machine.

The idea that such experiments can describe the capacity of a learning machine (it was called "effective VC dimension") was suggested in an article by Vapnik, Levin, and LeCun (1994). In experiments conducted by E. Levin, Y. LeCun, and later I. Guyon, C. Cortes, and P. Laskov with machines estimating linear functions, high precision of this method was shown.

It should be noted that the idea that for randomly labeled data the deviation of the minimum value of empirical risk from 1/2 can be used to define bounds for prediction error has been under discussion for more than 20 years. In the 1970s it was discussed by Pinsker (1979), and in the 1990s Brailovsky (1994) reintroduced this idea. The subject of analysis was the hypothesis that this deviation defines the value of confidence interval. However, as was shown in Chapter 4 the value of the confidence interval depends not only on the number of observations and on the capacity of the set of functions, but on the value of the empirical risk as well. It looks more realistic that expectations of the deviation define the capacity of the learning machine. Using the estimated capacity and the obtained bound (maybe with different constants), one can estimate the prediction accuracy. The method described in Chapter 13 is a step in this direction.

The main problem in approximation of data by a smooth function is to control the trade-off between accuracy of approximation of the data and complexity of the approximating function. In the experiments with approximation of the *sinc*-function by linear splines with an infinite number of knots, we demonstrated that by controlling the insensitivity value ε in the SV technique one can effectively control such a trade-off.

The problem of regression estimation is one of the key problems of applied statistics. Before the 1970s the main approach to estimating multi-dimensional regression was constructing linear approximating functions using more or less the same techniques, such as the least-squares method

or the least modulus method (robust methods). The set of linear functions, however, often turns out to be too poor to approximate the regression function well. Therefore in the 1970s generalized linear functions were introduced (set of functions that are a linear combination of a relatively small number of basis functions). Researchers hoped that they could define a reasonably small number of basis functions that would make it possible to approximate the unknown regression well. The experiments, however, showed that it is not easy to choose such a basis.

In 1980–1990 the natural generalization of this idea, the so-called dictionary method, was suggested. In this method, one defines a priori a large (possibly infinite) number of possible bases and uses training data both for choosing the small number of basis functions from the given set and for estimating the appropriate coefficients of expansion on a chosen basis. The dictionary method includes such methods as Projection Pursuit (see Friedman and Stuetzle (1981), Huber (1985)) and MARS (Multivariate Adaptive Regression Spline) (see Friedman (1991)). The last method is very attractive from both analytical and computational points of view and therefore became an important tool in multidimensional analysis.

In contrast to the dictionary method, the SV method suggests using all elements of the dictionary and controlling capacity by a special type of regularization. In other words, both the dictionary method and the SV method realize the SRM induction principle. However, they use different types of structures on the set of admissible functions. Moreover, both the MARS method and the SV method with kernels for constructing splines use the same dictionary containing tensor products of basis functions defined by polynomial splines. Therefore, comparison of MARS-type methods with the SV machine is, in fact, comparison of two different ideas of capacity control: by model selection and by regularization (assuming that both algorithms choose the best possible parameters).

The experiments described in this chapter demonstrate the advantage of regularization compared to feature selection.

The described method of solving linear equations is a straightforward generalization of the SV regression estimation method. However, it gives two new opportunities in solving the PET problem:

- One can exclude the pixel-parameterization of the solution.
- One can use a more sophisticated scheme of regularization by treating different measurements with different levels of accuracy.

The challenging problem in PET is to obtain a 3D solution. The main difficulty in solving 3D PET using the classical technique is the necessity of voxel-parameterization (3D piecewise constant functions). The 2D pixel representation contains a 256×256 constant to be estimated which is approximately equal to the number of observations. In the 3D problem, one has to estimate

$256 \times 256 \times 256$ parameters using observations which number much less than the number of parameters. The considered features of PET solution using the SV technique give a hope that 3D solutions are possible.

CHAPTERS 14, 15, AND 16

These chapters are written on the basis of articles by Vapnik and Chervonenkis (1971, 1981, and 1989).

EPILOGUE: INFERENCE FROM SPARSE DATA

Statistical learning theory does not belong to any specific branch of science: It has its own goals, its own paradigm, and its own techniques.

In spite of the fact that the first publications presented this theory as results in statistics, statisticians (who had their own paradigm) never considered this theory as a part of statistics.

Probabilists started using these techniques approximately 10 years after their introduction. They adopted the new ideas, reconsidered the Glivenko–Cantelli problem, developed the theory of the uniform central limit theorem, and obtained asymptotically sharp bounds on the uniform law of large numbers. However, they were not interested in developing inductive principles for function estimation problems.

In the mid 1980s, computer scientists tried to absorb part of this theory. In 1984 the probably approximately correct (PAC) model of learning was introduced (Valiant, 1984) which combined a simplified statistical learning model with an analysis of computational complexity. In this model, however, the statistical part was too simplified; it was restricted by problems of learnability. As was shown by Blumer et al. (1989) the constructions obtained in statistical learning theory give the complete answer to the PAC problem.

In the last few years mathematicians have become interested in learning theory. Two excellent books devoted to mathematical problems of learning appeared: *A Probabilistic Theory of Pattern Recognition* by L. Devroye, L. Gyorfi, and G. Lugosi (1996) and *A Theory of Learning and Generalization* by M. Vidyasagar (1997). In these books, the conceptual line of statistical learning theory is described with great art. However, another aspect of the problem exists: Using our new understanding of the models of generalization to construct efficient function estimation methods.

Statistical learning theory is such that all parts of it are essential: Any attempt to simplify it or separate one part from another harms the theory, its philosophy, and its methods for applications. In order to accomplish its goals, the theory should be considered as a whole.

Learning theory has one clear goal: to understand the phenomenon of induction that exists in nature. Pursuing this goal, statistical learning theory has

obtained results that have become important for many branches of mathematics and in particular for statistics. However, further study of this phenomenon requires analysis that goes beyond pure mathematical models.

As does any branch of natural science, learning theory has two sides:

1. The *mathematical* side that describes laws of generalization which are valid for all possible worlds and
2. The *physical* side that describes laws which are valid for our specific world, the world where we have to solve our applied tasks.

From the mathematical part of learning theory it follows that machines can generalize only because they use elements of a structure with restricted capacity. Therefore machines cannot solve the overwhelming majority of possible formal problems using small sample sizes. To be successful, learning machines must use structures on the set of functions that are appropriate for problems of our world. Ignoring this fact can lead to destructive analysis (as shown by the criticism of perceptrons in the late-1960s and the criticism of learning theories based on "no free lunch theorems" in the mid-1990s).

This book mostly considers the mathematical part of the problem. However, in solving applied problems we observed some phenomena that can be considered a raw material for analysis of physical laws of our world; the advantage of certain structures over others, the important role of invariants, the same support vectors for different kernels, etc.

Constructing the physical part of the theory and unifying it with the mathematical part should be considered as one of the main goals of statistical learning theory.

To achieve this goal we have to concentrate on the problem which can be called

Inference from Sparse Data

where, in order to generalize well, one has to use both mathematical and physical factors.

In spite of all results obtained, statistical learning theory is only in its infancy: There are many branches of this theory that have not yet been analyzed and that are important both for understanding the phenomenon of learning and for practical applications. They are waiting for their researchers.

REFERENCES

- S. Abou-Jaoude (1976), Conditions nécessaires et suffisantes de convergence L_1 en probabilité de l'histogramme pour une densité. *Ann. de Inst. H. Poincaré Ser. B* (12), 213–231.
- F. A. Aidu and V. N. Vapnik (1989), Estimation of probability density on the basis of the method of stochastic regularization, *Autom. Remote Control* 4, 84–97.
- M. A. Aizerman, E. M. Braverman, and L. I. Rozonoer (1964a), Theoretical foundations of the potential function method in pattern recognition learning, *Autom. Remote Control* 25, 821–837.
- M. A. Aizerman, E. M. Braverman, and L. I. Rozonoer (1964b), The problem of pattern recognition learning and the method of potential functions, *Autom. Remote Control* 25, 1175–1193.
- M. A. Aizerman, E. M. Braverman, and L. I. Rozonoer (1970), *Method of Potential Functions in the Theory of Pattern Recognition* (in Russian), Nauka, Moscow, p. 384.
- H. Akaike (1970), Statistical predictor identification, *Ann. Inst. Stat. Math.*.. 202–217.
- K. Alexander (1984), Probability inequalities for empirical processes and law of iterated logarithm, *Ann. Probab.* 4, 1041–1067.
- D. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler (1993). Scale-sensitive dimensions, uniform convergence, and learnability, in *Proceedings of the 34th Annual IEEE Conference on Foundations of Computer Science.*, pp. 292–301.
- S. Amari (1967), A theory of adaptive pattern classifiers, *IEEE Trans. Electron Comput.* EC-16, 299–307.
- T. W. Anderson and R. R. Bahadur (1966), Classification into two multivariate normal distributions with different covariance matrices, *Ann. Math. Stat.* **133**(2).
- N. Aronszajn (1943), The theory of reproducing kernels and their applications, *Cambridge Phil. Soc. Proc.* 39, 133–153.
- A. R. Barron (1993), Universal approximation bounds for superpositions of a sigmoid function, *IEEE Trans. Inf. Theory* **39**(3), 930–945.
- A. R. Barron and T. Cover (1991), Minimum complexity density estimation. *IEEE Trans. Inf. Theory* **37**, 1034–1054.
- P. Bartlett (1997), For valid generalization the size of the weights is more important than the size of network, in *Advances in Neural Information Processing Systems*, Vol. 9, MIT Press, Cambridge, MA, pp. 134–140.

- D. Belsley, E. Kuh, and R. Welsch (1980), *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, Wiley, New York.
- J. Berger (1985). *Statistical Decision Theory and Bayesian Analysis*, Springer, New York.
- A. Blumer, A. Ehrenfeucht, D. Haussler and M. Warmuth (1989), Learnability and the Vapnik–Chervonenkis dimension, *J. ACM* **36**, 929–965.
- S. Bochner (1932), *Vorlesungen über Fourierche Integrale*, Academisch Verlagsgesellschaft, Leipzig. English translation: S. Bochner (1959) *Lectures on Fourier Integral*, *Ann. Math. Stud.*, 42.
- B. Boser, I. Guyon, and V. N. Vapnik (1992), A training algorithm for optimal margin classifiers, in *Fifth Annual Workshop on Computational Learning Theory*. ACM, Pittsburgh, pp. 144–152.
- L. Bottou and V. Vapnik (1992), Local learning algorithms, *Neural Comput.* **4**(6), pp. 888–901.
- L. Bottou, C. Cortes, J. Denker, H. Drucker, I. Guyon, L. Jackel, Y. LeCun, U. Müller, E. Sackinger, P. Simard, and V. Vapnik (1994), Comparison of classifier methods: A case study in handwritten digit recognition, in *Proceedings of the 12th IAPR International Conference on Pattern Recognition*, Vol. 2, IEEE Computer Society Press, Los Alamitos, CA, pp. 77–83.
- V. L. Brailovsky (1994). Probabilistic approach to model selection: Comparison with unstructured data, in *Selecting Models from Data: AI and Statistics*, Springer-Verlag, New York, pp. 61–70.
- L. Breiman (1993). Hinging hyperplanes for regression, classification and function approximation, *IEEE Trans. Inf. Theory* **39**(3), 999–1013.
- L. Breiman (1996), Bagging predictors, *Mach. Learn.* **24**(2), 123–140.
- L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone (1984). *Classification and Regression Trees*, Wadsworth, Belmont, CA.
- A. Bruce, D. Donoho, and H. Y. Gao (1996), Wavelet analysis, *IEEE Spectrum* Oct. 26–35.
- A. Bryson, W. Denham, and S. Dreyfus (1963), Optimal programming problem with inequality constraints. I: Necessary conditions for extremal solutions, *AIAA J.* **1**, 2544–2550.
- J. Bunch, L. Kaufman (1980), A computational method for the indefinite quadratic optimization problem, *Linear Algebra and Its Applications* **34**, 341–370.
- C. Burges (1996), Simplified support vector decision rules, in *Proceedings of the ICML '96*, Bari, Italy.
- C. Burges and B. Scholkopf (1997), Improving the accuracy and speed of support vector machines, in *Advances in Neural Information Processing Systems*, Vol. 9, The MIT Press, Cambridge, MA.
- F. P. Cantelli (1933), Sulla determinazione empirica della leggi di probabilità, *C. Inst. Ital. Attiari* **4**.
- G. J. Chaitin (1966), On the length of programs for computing finite binary sequences, *J. Assoc. Comput. Mach.* **13**, 547–569.
- S. Chen, D. Donoho, and M. Saunders (1995), "Atomic Decomposition by Basis Pursuit." Technical Report 479, Department of Statistics, Stanford University.

- N. N. Chentsov (1963), Evaluation of an unknown distribution density from observations, *Soviet Math.* 4, 1559–1562.
- N. N. Chentsov (1981), On correctness of problem of statistical point estimation. *Theory Probab. Appl.*, **26**, 13–29.
- V. Cherkassky and F. Mulier (1998), *Learning from Data: Concepts, Theory, and Methods*, Wiley, New York.
- V. Cherkassky, F. Mulier, and V. Vapnik (1996), Comparison of VC method with classical methods for model selection, in *Proceedings of the World Congress on Neural Networks*, San Diego, CA, pp. 957–962.
- H. Chernoff (1952), A measure of asymptotic efficiency of test of a hypothesis based on the sum of observations, *Ann. Math. Stat.* 23, 493–507.
- C. Cortes and V. Vapnik (1995), Support vector networks, *Mach. Learn.* 20, 1–25.
- R. Courant and D. Hilbert (1953), *Methods of Mathematical Physics*, Wiley, New York.
- H. Cramer (1946), *Mathematical Methods of Statistics*, Princeton University Press, Princeton, NJ.
- P. Craven and H. Wahba (1979), Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation, *Numer. Math.* 31, 377–403.
- G. Cybenko (1989), Approximation by superpositions of sigmoidal function. *Math. Control Signals Syst.* 2, 303–314.
- L. Devroye (1988), Automatic pattern recognition: A study of the probability of error. *IEEE Trans. Pattern Anal. Mach. Intell.* **10**(4), 530–543.
- L. Devroye and L. Gyorfi (1985), *Nonparametric Density Estimation: The L_1 View*, Wiley, New York.
- L. Devroye, L. Györfi, and G. Lugosi (1996), *A Probabilistic Theory of Pattern Recognition*, Springer, New York.
- P. Diaconis and D. Freedman (1986), On the consistency of Bayesian Estimates (with discussions), *Ann. Stat.* **14**(1), 1–67.
- H. Drucker (1997), Improving regression using boosting techniques, in *Proceedings of the International Conference on Machine Learning* (ICML '97). D. H. Fisher Jr., ed., Morgan Kaufmann, San Mateo, CA, pp. 107–113.
- H. Drucker, C. Burges, L. Kaufman, A. Smola, and V. Vapnik (1997), Support vector regression machines, in *Advances in Neural Information Processing Systems*, Vol. 9, MIT Press, Cambridge, MA.
- H. Drucker, R. Schapire, and P. Simard (1993), Boosting performance in neural networks, *Int. J. Pattern Recognition Artif. Intell.* **7**(4), 705–719.
- R. M. Dudley (1978), Central limit theorems for empirical measures, *Ann. Probab.* **6**(6), 899–929.
- R. M. Dudley (1984), *Course on Empirical Processes*, Lecture Notes in Mathematics, Vol. 1097, Springer, New York, pp. 2–142.
- R. M. Dudley (1987), Universal Donsker classes and metric entropy, *Ann. Probab.* **15**(4), 1306–1326.

- A. Dvoretzky, J. Kiefer, and J. Wolfowitz (1956), Asymptotic minimax character of a sample distribution function and of the classical multinomial estimator, *Ann. Math. Stat.* 33, pp. 642–669.
- H. W. Engl, M. Hanke, and A. Neubauer (1996), *Regularization of Inverse Problems*, Kluwer Academic Publishers, Hingham, MA, 318 pages.
- R. A. Fisher (1952), *Contributions to Mathematical Statistics*, Wiley, New York.
- Y. Freund and R. Schapire (1995), A decision-theoretic generalization of on-line learning and an application to learning, in *Computational Learning Theory*, Springer, New York, pp. 23–37.
- V. Fridman (1956), Methods of successive approximations for Fredholm integral equation of the first kind, *Uspekhi Math. Nauk* 11, 1 (in Russian).
- J. H. Friedman (1991), Multivariate adaptive regression splines, *Ann. Stat.* (with discussion) 19, 1–141.
- J. H. Friedman and W. Stuetzle (1981). Projection pursuit regression, *JASA* 76, 817–823.
- E. Giné and J. Zinn (1984). Some limit theorems for empirical processes. *Ann. Probab.* 12(4), 929–989.
- F. Girosi (1998), An equivalence between sparse approximation and Support Vector Machines. *Neural Computation* (to appear).
- F. Girosi, Approximation error bounds that use VC bounds. in *Proceedings of ICANN '95*, Paris, 1995.
- F. Girosi and G. Anzellotti (1993). Rate of convergence for radial basis functions and neural networks. *Artificial Neural Networks for Speech and Vision*, Chapman & Hall, London, pp. 97–113.
- V. I. Glivenko (1933), Sulla determinazione empirica di probabilità, *G. Inst. Ital. Attuari* 4.
- I. S. Gradshteyn and I. M. Ryzhik (1980), *Table of Integrals, Series, and Products*, Academic Press, New York.
- U. Grenander (1981), *Abstract Inference*, Wiley, New York.
- J. Hadamard (1902). Sur les problèmes aux dérivées partielles et leur signification physique, *Bull. Univ. Princeton* 13, 49–52.
- D. Haussler (1992). Decision theoretic generalization of the PAC model for neural nets and other learning applications, *Inf. Comput.* 100, 78–150.
- T. J. Hastie and R. J. Tibshirani (1990). *Generalized Linear Models*, Chapman and Hall, London.
- A. E. Hoerl and R. W. Kennard (1970), Ridge regression: Biased estimation for non-orthogonal problems, *Technometrics* 12, 55–67.
- P. Huber (1964), Robust estimation of location parameter, *Ann. Math. Stat.*, 35(1).
- P. Huber (1985), Projection pursuit, *Ann. Stat.* 13, 435–475.
- W. Hrdlíc (1990). *Applied Nonparametric Regression*. Cambridge: Cambridge University Press.
- I. A. Ibragimov and R. Z. Hasminskii (1981), *Statistical Estimation: Asymptotic Theory*, Springer, New York.
- V. V. Ivanov (1962). On linear problems which are not well-posed, *Soviet Math. Dokl.* 3(4), 981–983.

- V. V. Ivanov (1976). *The Theory of Approximate Methods and Their Application to the Numerical Solution of Singular Integral Equations*, Nordhoff International, Leyden.
- W. James and C. Stein (1961), Estimation with quadratic loss, in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, University of California Press, Berkeley, CA.
- Th. Joachims (1998) Making large-scale SVM learning practical. In: B. Scholkopf, C. Burges, A. Smola (eds.) *Advances in Kernel methods—Support Vector Learning*. MIT Press, Cambridge, MA.
- L. K. Jones (1992). A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression, *Ann. Stat.* 20(1). 608–613.
- M. Karpinski and T. Werther (1989), VC dimension and uniform learnability of sparse polynomials and rational functions, *SIAM J. Comput.* (Preprint 8537-CS. Bonn University, 1989.)
- M. Kearns and R. Schapire (1994), Efficient distribution-free learning of probabilistic concepts, *J. Computer and System Sci.* 48(3), 464–497.
- V. I. Kolchinskii (1981), On central limit theorem for empirical measures, *Theory of Probability and Mathematical Statistics* 2.
- A. N. Kolmogorov (1933a), Sulla determinazione empirica di una legge di distribuzione, *G. Inst. Ital. Attuari* 4.
- A. N. Kolmogorov (1933b), *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Springer, Berlin. (English translation: A. N. Kolmogorov (1956), *Foundation of the Theory of Probability*, Chelsea, New York.)
- A. N. Kolmogorov (1965), Three approaches to the quantitative definitions of information. *Prob. Inf. Transm.* 1(1), 1–7.
- A. N. Kolmogorov and S. V. Fomin (1970), *Introductory Real Analysis*, Prentice-Hall. Englewood Cliffs, NJ.
- M. M. Lavrentiev (1962), *On Ill-Posed Problems of Mathematical Physics*, Novosibirsk, SO AN SSSR (in Russian).
- L. LeCam (1953), On some asymptotic properties of maximum likelihood estimates and related Bayes estimates, *Univ. Calif. Public Stat.* 11.
- Y. LeCun (1986), Learning processes in an asymmetric threshold network, *Disordered Systems and Biological Organizations*, Springer, Les Houches, France, pp. 233–240.
- Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. J. Jackel (1990), Handwritten digit recognition with back-propagation network, in *Advances in Neural Information Processing Systems*, Vol. 2, Morgan Kaufman, San Mateo, CA, pp. 396–404.
- Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner (1998), Gradient-based learning applied to document recognition, *Proceedings of the IEEE*. Special Issue on Intelligent Signal Processing.
- G. G. Lorentz (1966), *Approximation of Functions*, Holt-Rinehart-Winston, New York.
- A. Luntz and V. Brailovsky (1969), On estimation of characters obtained in statistical procedure of recognition, (in Russian) *Technicheskaya Kibernetika*. 3.
- N. M. Markovich (1989), Experimental analysis of non-parametric estimates of a probability density and methods for smoothing them. *Autom. and Remote Contr.* 7.

- P. Massart (1990), The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality, *Ann. Probab.* 18, 1269–1283.
- T. Mercer (1909), Functions of positive and negative type and their connection with the theory of integral equations, *Trans. Lond. Philos. Soc. A*, 209, 415–446.
- H. N. Mhaskar (1993), Approximation properties of a multi-layer feed-forward artificial neural network, *Adv. Comput. Math.* 1, 61–80.
- C. A. Micchelli (1986), Interpolation of scattered data: Distance matrices and conditionally positive definite functions, *Constr. Approx.* 2, 11–22.
- S. G. Mikhlin (1964), *Variational Methods of Mathematical Physics*. Pergamon Press, Oxford.
- M. L. Miller (1990), *Subset Selection in Regression*, Chapman & Hall, London.
- M. L. Minsky and S. A. Papert (1969), *Perceptrons*, MIT Press, Cambridge, MA.
- J. E. Moody (1992), The effective number of parameters: An analysis of generalization and regularization in non-linear learning systems, in *Advances in Neural Information Processing Systems*, Vol. 5, Morgan Kaufmann, San Mateo, CA.
- J. J. More and G. Toraldo (1991), On the solution of large quadratic programming problems with bound constraints, *SIAM Optim.* 1(1), 93–113.
- S. Mukherjee, E. Osuna, and F. Girosi (1997), Nonlinear prediction of chaotic time series using support vector machines, in *Proceedings of IEEE Conference Neural Networks for Signal Processing*, Amelia Island.
- K. R. Müller, A. J. Smola, G. Ratsch, B. Scholkopf, J. Kohlomorgen, and V. Vapnik (1997), Predicting time series with support vector machines, in *Proceedings of the 1997 ICANN Conference*.
- B. Murtagh and M. Saunders (1978), Large-scale linearly constrained optimization, *Math. Program.* 14, 41–72.
- A. B. J. Novikoff (1962), On convergence proofs on perceptrons, in *Proceedings of the Symposium on the Mathematical Theory of Automata*, Vol. XII, Polytechnic Institute of Brooklyn, pp. 615–622.
- E. Osuna, R. Freund, and F. Girosi (1997a), Training support vector machines: An application to face detection, in *Proceedings 1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Computer Society, Los Alamos.
- E. Osuna, R. Freund, and F. Girosi (1997b), Improved training algorithm for support vector machines, in *Proceedings of IEEE Conference Neural Networks for Signal Processing*, Amelia Island.
- J. M. Parrondo and C. Van den Broeck (1993), Vapnik–Chervonenkis bounds for generalization, *J. Phys. A* 26, 2211–2223.
- E. Parzen (1962), On estimation of probability function and mode, *Ann. Math. Stat.* 33(3).
- D. Z. Phillips (1962), A technique for numerical solution of certain integral equation of the first kind, *J. Assoc. Comput. Mach.* 9, 84–96.
- I. I. Pinsker (1979), The chaotization principle and its application in data analysis (in Russian), in *Models, Algorithms, Decision Making*, Nauka, Moscow.
- J. Platt (1998), Sequential minimal optimization: A fast algorithm for training support vector machines. In: B. Scholkopf, C. Burges, A. Smola (eds.) *Advances in Kernel Methods—Support Vector Learning*. MIT Press, Cambridge, MA.

- T. Poggio and F. Girosi (1990), Networks for approximation and learning, Proc. IEEE, **78**(9).
- D. Pollard (1984), Convergence of Stochastic Processes, Springer, New York.
- K. Popper (1968), The Logic of *Scientific* Discovery, 2nd ed., Harper Torch Book, New York.
- M. J. D. Powell (1992), The theory of radial basis functions approximation in 1990. Advances in Numerical Analysis Volume *II*: Wavelets, Subdivision Algorithms and Radial Basis Functions, W. A. Light, ed., Oxford University, pp. 105–210.
- J. Rissanen (1978), Modeling by shortest data description, *Autontatica* **14**, 465–471.
- J. Rissanen (1989), Stochastic Complexity and Statistical Inquiry, World Scientific.
- H. Robbins and S. Monroe (1951), A stochastic approximation method, *Ann. Math. Stat.* 22, 400–407.
- F. Rosenblatt (1962), Principles of Neurodynamics: Perceptron and Theory of *Brain* Mechanisms, Spartan Books, Washington, DC.
- M. Rosenblatt (1956), Remarks on some nonparametric estimation of density function, *Ann. Math. Stat.* 27, 642–669.
- D. E. Rumelhart, G. E. Hinton, and R. J. Williams (1986), Learning internal representations by error propagation, Parallel *Distributed* Processing: Explorations in the Macrostructure of Cognition, Vol. I, Bradford Books, Cambridge, MA, pp. 318–362.
- N. Sauer (1972), On the density of families of sets, *J. Cotnb. Theory (A)* **13**, 145–147.
- M. Saunders and B. Murtagh (1994), MINOS 5.4 User's Guide, Report SOL 83-20R, Department of Operations Research, Stanford University (revised Feb. 1995).
- R. E. Schapire, Y. Freund, P. Bartlett, and W. Sun Lee (1997), Boosting the margin: A new explanation for the effectiveness of voting methods, in Machine Learning: Proceedings of the Fourteenth International Conference, pp. 322–330.
- B. Scholkopf, A. Smola, and K. Miiller (1997), Kernel principal component analysis, in ICANN '97.
- B. Scholkopf, P. Simard, A. Smola, and V. Vapnik (1998), Prior knowledge in support vector kernels, in Advances in Neural Information Processing Systems, Vol. 10, MIT Press, Cambridge, MA.
- G. Schwartz (1978), Estimating the dimension of a model, *Ann. Stat.* **6**, 461–464.
- J. Shao (1993), Linear model selection by cross-validation, *J. Am. Stat. Assoc. Theory and Methods* **422**.
- S. Shelah (1972), A combinatorial problem: Stability and order of models and theory of infinitary languages, *Pacific J. Math.* **41**, 247–261.
- R. Shibata (1981), An optimal selection of regression variables, *Biometrika* **68**, 461–464.
- A. N. Shirayev (1984), Probability, Springer, New York.
- P. Y. Simard, Y. LeCun, and J. Denker (1993), Efficient pattern recognition using a new transformation distance, *Neural Inf. Processing Syst.* **5**, 50–58.
- N. V. Smirnov (1970), Theory of Probability and Mathematical Statistics (Selected *Works*), Nauka, Moscow (in Russian).
- R. J. Solomonoff (1960), A Preliminary Report on General Theory of Inductive Inference, Technical Report ZTB-138, Zator Company, Cambridge, MA.

- R. J. Solomonoff (1964), A formal theory of inductive inference. Parts 1 and 2, *Inf. Control* 7, 1–22. 224–254.
- A. R. Stcfanyuk (1986), Estimation of the likelihood ratio function in the “disorder” problem of random processes, *Autom. Remote Control* 9, 53–59.
- J. M. Steele (1978), Empirical discrepancies and subadditive processes, *Ann. Probab.* 6, 118–127.
- E. M. Stein (1970), *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, NJ.
- J. Stewart (1976), Positive definite functions and generalizations, historical survey. *Rocky Mountain J. Math.* 6(3), 409–434.
- C. Stone, M. Hansen, C. Kooperberg, and Y. Throung (1997), Polynomial splines and their tensor products in extended linear modeling (with discussion), *Ann. Stat.* 25 (to appear).
- W. Stute (1986), On almost sure convergence of conditional empirical distribution function, *Ann. Probab.* 14(3), 891–901.
- M. Talegrand (1994). Sharper bounds for Gaussian and empirical processes, *Ann. Probab.* 22.
- R. A. Tapia and J. R. Thompson (1978). *Nonparametric Probability Density Estimation*, Johns Hopkins University Press, Baltimore.
- A. N. Tikhonov (1943), On the stability of inverse problem, *Dokl. Acad. Nauk USSR* 39, 5 (in Russian).
- A. N. Tikhonov (1963). On solving ill-posed problem and method of regularization. *Dokl. Akad. Nauk USSR* 153, 501–504.
- A. N. Tikhonov and V. Y. Arsenin (1977), *Solution of Ill-Posed Problems*, W. H. Winston, Washington, DC.
- E. C. Titchmarsh (1948), *Introduction to Theory of Fourier Integrals*. The Clarendon Press, Oxford.
- Ya. Z. Tsypkin (1971), *Adaptation and Learning in Automatic Systems*, Academic Press, New York.
- Ya. Z. Tsypkin (1973). *Foundation of the Theory of Learning Systems*, Academic Press. New York.
- M. Unser and A. Aldroubi (1992), Polynomial splines and wavelets, in *Wavelets — A Tutorial in Theory and Applications*, C. K Chui, ed.. pp. 91–122.
- L. Valiant (1984), A theory of the learnability. *Commun. ACM* 27, 1134–1142.
- R. Vanderbei (1994), LOQO: An Interior Point Code for Quadratic Programming, Technical Report SOR-94-15.
- A. W. van der Vaart and J. A. Wellner (1996), *Weak Convergence and Empirical Processes*, Springer. New York
- V. N. Vapnik (1979). *Estimation of Dependencies Based on Empirical Data*, Nauka, Moscow (in Russian). (English translation: V. Vapnik (1982), *Estimation of Dependencies Based on Empirical Data*, Springer, New York.)

- V. N. Vapnik (1988), Inductive principles of statistics and learning theory, in *Yearbook of the Academy of Sciences of the USSR on Recognition, Classification, and Forecasting* (in Russian), Vol. 1, Nauka, Moscow. (English translation: VN. Vapnik (1995), Inductive principles of statistics and learning theory, in *Mathematical Perspectives on Neural Networks*, Smolensky, Mozer, and Rumelhart, eds., Lawrence Erlbaum Associates.
- V. N. Vapnik (1993), Three fundamental concepts of the capacity of learning machines, *Physica A* **200**, 538–544.
- V. N. Vapnik (1995), *The Nature of Statistical Learning Theory*, Springer, New York.
- V. N. Vapnik and L. Bottou (1993), Local Algorithms for Pattern Recognition and Dependencies Estimation, *Neural Comput.* **5**(6), 893–908.
- V. N. Vapnik and A. Ya. Chervonenkis (1964), On one class of perceptrons, *Autom. and Remote Contr.* **25**(1).
- V. N. Vapnik and A. Ya. Chervonenkis (1968), On the uniform convergence of relative frequencies of events to their probabilities, *Soviet Math. Dokl.* **9**, 915–918.
- V. N. Vapnik and A. Ya. Chervonenkis (1971), On the uniform convergence of relative frequencies of events to their probabilities," *Theory Probab. Appl.* **16**, 264–280.
- V. N. Vapnik and A. Ya. Chervonenkis (1974), *Theory of Pattern Recognition*, Nauka, Moscow (in Russian). (German translation: W. N. Wapnik, A. Ya. Tscherwonenkis (1979), *Theorie der Zeichenerkennung*, Akademie, Berlin.)
- V. N. Vapnik and A. Ya. Chervonenkis (1981), Necessary and sufficient conditions for the uniform convergence of the means to their expectations, *Theory Probab. Appl.* **26**, 532–553.
- V. N. Vapnik and A. Ya. Chervonenkis (1989), The necessary and sufficient conditions for consistency of the method of empirical risk minimization, *Yearbook of the Academy of Sciences of the USSR*, on Recognition, Classification, and Forecasting, Vol. 2, Nauka, Moscow, pp. 207–249 (in Russian). (English translation: V. N. Vapnik and A. Ya. Chervonenkis (1991), The necessary and sufficient conditions for consistency of the method of empirical risk minimization, *Pattern Recogn. Image Anal.* **1**(3), 284–305.)
- V. N. Vapnik, S. E. Golowich, and A. Smola (1997), Support vector method for function approximation, regression, and signal processing, in *Advances in Neural Information Processing Systems*, Vol. 9, MIT Press, Cambridge, MA.
- V. Vapnik and A. Lerner (1963) Pattern recognition using generalized portrait method. *Autom. Remote Control*. 24.
- V. Vapnik, E. Levin, and Y. LeCun (1994), Measuring the VC Dimension of a Learning Machine, *Neural Computation* **10**(5).
- V. N. Vapnik, N. M. Markovich, and A. R. Stefanyuk (1992). Rate of convergence in L_2 of the projection estimator of the distribution density, *Autom. Remote Control* 5.
- V. N. Vapnik and A. R. Stefanyuk (1978), Nonparametric methods for estimating probability densities, *Autom. Remote Control* 8.
- V. N. Vapnik and A. M. Sterin (1977), On structural risk minimization of overall risk in a problem of pattern recognition, *Autom. Remote Control* 10.
- M. Vidyasagar (1997), *A Theory of Learning and Generalization with Application to Neural Networks and Control Systems*, Springer, New York.

- V. Vovk (1992), Universal forecasting algorithms, *Information and Computation* **96**(2) 245–277.
- B. von Bahr and G. G. Essen (1965), Inequalities for r th absolute moment of a sum of random variables, $1 < r \leq 2$, *Ann. Math. Stat.* **36**(1), 299–303.
- G. Wahba (1990), Spline Model *for* Observational Data, Society for Industrial and Applied Mathematics, Philadelphia.
- C. S. Wallace and D. M. Boulton (1968), An information measure for classification, *Comput. J.* **11**, 185–195.
- G. Watson (1964), Smooth regression analysis, *Sankhya, Series A* (26), 359–372.
- R.S. Wenocur and R. M. Dudley (1981), Some special Vapnik–Chervonenkis classes, *Discrete Math.* **33**, 313–318.
- J. Weston, A. Gammerman, M. Stitson, V. Vapnik and V. Vovk (1998) Density estimator using support vector machines. In: B. Scholkopf, C. Burges, A. Smola (eds.) *Advances in Kernel Methods—Support Vector Learning*. MIT Press, Cambridge. MA.
- D. H. Wolpert (1995), The relationship between PAC, the statistical physics framework, the Bayesian framework, and the VC framework, in *Mathematics of Generalization, Proceedings Santa Fe Institute Studies of Complexity*, Volume XX. Addison-Wesley Publishing Company.